EXPLORING HUMAN POPULATION VARIATION

IN REPEAT PROTEINS

By

Maxim Ivanov Tsenkov

Thesis submitted for the degree of Doctor of Philosophy

at

University of Dundee

Dundee, United Kingdom

October 2022

Copyright by Maxim Tsenkov, 2022
Contents

List of Figures ........................................................................................................ iv
List of Tables ........................................................................................................ vi
List of Abbreviations ............................................................................................ vii
Acknowledgements ............................................................................................... x
STATEMENT .......................................................................................................... xi
DECLARATION ....................................................................................................... xii
Abstract ................................................................................................................ xiii

Introduction............................................................................................................. 1

1.1 Preface .............................................................................................................. 1
1.2 Background ...................................................................................................... 1

1.3 The Era of Big Data ......................................................................................... 2
  1.3.1 Functional annotation of protein sequences .............................................. 4
  1.3.2 Predictive diagnostic models ..................................................................... 4
  1.3.3 Multiple sequence alignments ................................................................. 7
  1.3.4 The Protein Data Bank ............................................................................. 10
  1.3.5 Structure classification ............................................................................ 11
  1.3.6 Protein domain architecture .................................................................... 11
  1.3.7 Methods for structure comparison ........................................................... 12
  1.3.8 The most common classes of protein architectures .................................. 12

1.4 Repeat families ................................................................................................ 16
  1.4.1 Solenoid proteins ....................................................................................... 16
  1.4.2 Topological arrangements of helical repeat families ............................... 16
  1.4.3 Structural features of solenoid proteins ................................................. 18
  1.4.4 Protein–protein interactions .................................................................... 20
  1.4.5 Repeat sequences ...................................................................................... 25
  1.4.6 The methodological challenges for protein repeat detection .................. 29

1.5 Genetic variation ............................................................................................. 30
  1.5.1 Types genetic variation ............................................................................ 32
  1.5.2 Evolution of protein-coding regions in the genome ............................... 33
  1.5.3 Interpreting variants between individual protein residues ..................... 34

1.6 Scope of the thesis ......................................................................................... 36

Development of RAM ........................................................................................... 39

2.1 Preface .............................................................................................................. 39
2.2 Introduction ..................................................................................................... 40

2.3 Methods and Contents .................................................................................... 41
  2.3.1 Gathering repeat annotations ................................................................... 43
  2.3.2 Repeat architecture ................................................................................... 51
  2.3.3 Multiple sequence alignment ................................................................... 62
  2.3.4 Evaluating the quality of a multiple sequence alignment ....................... 67
  2.3.5 Multiple sequence alignment analysis ..................................................... 73
List of Figures

FIGURE 1.1: SCHEMATIC REPRESENTATION OF THE MEMBER DATABASES MAKING UP INTERPRO ..................................5
FIGURE 1.2: SEQUENCE AND STRUCTURE DIVERSITY OF THE PERTACTIN B-SOLENOID STRUCTURE ..................26
FIGURE 1.3: SEQUENCE LOGO AND CONSERVATION PROFILE OF REPEAT FAMILIES ..................................27
FIGURE 1.4: COST OF SEQUENCING HUMAN GENOMES ......................................................................................31
FIGURE 2.1: SCHEMATIC REPRESENTATION OF THE RAM PROTOCOL .................................................................42
FIGURE 2.2: ENTRY MATCHES FOR THREE REPEAT-CONTAINING PROTEINS IN INTERPRO ............................45
FIGURE 2.3: GROUPING SIMILAR REPEAT ANNOTATIONS WITH TPRs .................................................................49
FIGURE 2.4: SCHEMATIC REPRESENTATION OF REPEAT ARCHITECTURE DEMONSTRATED WITH THE TPR FAMILY. 54
FIGURE 2.5: REPEAT ARCHITECTURE ANALYSIS FOR THE TPR FAMILY ..............................................................56
FIGURE 2.6: STRUCTURAL ALIGNMENT OF ARMIADILLO REPEATS AND MSA ..........................................................58
FIGURE 2.7: ALPHA FOLD MODEL OF TETRATRICOPETIDE REPEAT PROTEIN 7B ..............................................61
FIGURE 2.8: RECURSIVE ALIGNMENT ALGORITHM .................................................................................................63
FIGURE 2.9: TEMPLATE MSA AND OUTPUT MSA FROM AMPS FOR ARMIADILLO REPEATS .................................66
FIGURE 2.10: FOUR MSAS BUILT USING DIFFERENT ALIGNMENT METHODS ........................................................70
FIGURE 2.11: EXTRACTS OF EACH TPR ALIGNMENT IN HOMO SAPIENS SPLIT BY TPR POSITION ..................78
FIGURE 2.12: THREE TPR SUBFAMILY MSAS WITH EACH MSA MAKING UP A TPR IN A 3-TPR-DOMAIN ...........82
FIGURE 2.13: SCHEMATIC REPRESENTATION OF MAXHMM ..................................................................................86
FIGURE 2.14: Z-SCORE HISTOGRAM FOR PAIRWISE SEQUENCE COMPARISONS OF THE TPR MSA .................87
FIGURE 2.15: PROCESSING HITS FROM THE HMMSEARCH TOOL .........................................................................89
FIGURE 2.16: OVERVIEW OF THE VARALIGN PIPELINE ........................................................................................93
FIGURE 2.17: SCHEMATIC REPRESENTATION OF MAPPING AND AGGREGATING VARIANTS ............................94
FIGURE 2.18: MISENSE ENRICHMENT PLOT ...........................................................................................................96
FIGURE 2.19: SUBSTITUTION MATRICES FOR THE ARMIADILLO MSA ..............................................................97
FIGURE 2.20: BUILDING CONTACT MAPS .................................................................................................................99
FIGURE 2.21: SECONDARY STRUCTURE ASSIGNMENT AND RSA CLASSIFICATIONS FOR THE TPR FAMILY ........102
FIGURE 3.1: TPR SEQUENCE LOGO AND STRUCTURAL FEATURES ....................................................................110
FIGURE 3.2: COMPARISON OF TPR ANNOTATIONS FROM FOUR DATABASES ..................................................114
FIGURE 3.3: REANNOTATING HUMAN TPRS WITH MAXHMM ...........................................................................115
FIGURE 3.4: PERFORMANCE OF A SINGLE HMM PROFILE AGAINST MAXHMM (MULTI-LEVEL HMM LIBRARY). 118
FIGURE 3.5: TPR MSA .............................................................................................................................................126
FIGURE 3.6: EVOLUTIONARY CONSERVED POSITIONS IN TPRS ........................................................................133
FIGURE 3.7: MISENSE ENRICHMENT PLOTS FOR TPR MSAS SPLIT BY REPEAT ARCHITECTURE ..................138
FIGURE 3.8: STRUCTURAL ANALYSIS OF INTRA- AND INTER- TPR CONTACTS ................................................145
FIGURE 3.9: MEDIAN RSA FOR EACH ALIGNMENT POSITION ACROSS START, MIDDLE, AND END TPRs ..........147
FIGURE 3.10: AGGREGATED TPR SUBSTRATE INTERACTIONS .........................................................................150
FIGURE 3.11: CLINVAR VARIANTS IN TPRS ...........................................................................................................153
FIGURE 4.1: ANKYrin REPEAT SEQUENCE AND STRUCTURE ..............................................................................158
FIGURE 4.2: COMPOSITION OF THE ANKYrin REPEAT ANNOTATION DATASET ..............................................160
FIGURE 4.3: OVERVIEW OF THE ANKYrin MSA ..................................................................................................162
FIGURE 4.4: NORMALISED SHENKIN CONSERVATION PROFILE ........................................................................164
FIGURE 4.5: MISENSE ENRICHMENT PLOT ...........................................................................................................165
FIGURE 4.6: INTRA-ANKYrin CONTACT ANALYSIS. (A) .......................................................................................168
FIGURE 4.7: MEDIAN RESIDUE SURFACE ACCESSIBILITY ....................................................................................169
FIGURE 4.8: REDEFINED SURFACES OF ANKYrin-DOMAINS ............................................................................170
FIGURE 4.9: MES ON ANKYrin-DOMAIN SURFACES ........................................................................................172
FIGURE 4.10: PROTEIN-SUBSTRATE INTERACTIONS ..........................................................................................173
FIGURE 4.11: SUBSTRATE INTERACTIONS WITH ANKYrin-DOMAINS .............................................................175
FIGURE 4.12: AGGREGATED SSE CLASSIFICATIONS. STACKED BAR PLOT OF THE SUMMARISED ................176
FIGURE 4.13: ASX MOTIFS IN THE ANKYrin REPEAT LOOP ..............................................................................178
FIGURE 4.14: CLINVAR VARIANTS IN THE ANKYrin REPEAT FAMILY ..............................................................179
FIGURE 5.1: ARMIADILLO REPEAT SEQUENCE LOGO AND STRUCTURE ........................................................186
FIGURE 5.2: FOUR EXAMPLES OF ARMIADILLO-DOMAINS AND THEIR INTERACTING PROTEIN SUBSTRATES 188
FIGURE 5.3: COMPARISON OF ANNOTATIONS FROM THREE DATABASES ............................................................191
FIGURE 5.4: ARMIADILLO MSA ............................................................................................................................193
FIGURE 5.5: TWO ADJACENT ARMADILLOS HIGHLIGHTING THREE AMINO ACID CLASSES OF CONSERVED POSITIONS. ................................................................. 196
FIGURE 5.6: SEQUENCE AND VARIANT CLASSIFICATIONS. ................................................................. 197
FIGURE 5.7: SUBSTITUTION MATRICES FOR VARIANTS AT FOUR POSITIONS IN THE ARMADILLO MSA .......... 199
FIGURE 5.8: STRUCTURAL ANALYSIS OF INTRA- AND INTER-ARMADILLO CONTACTS. ..................... 203
FIGURE 5.9: AGGREGATED SUBSTRATE INTERACTIONS. ........................................................................ 211
List of Tables

TABLE 1.1: CONTENTS AND COVERAGE OF THE INTERPRO PROTEIN MATCHES FOR ALL UNIPROT KB AND UNIPARC PROTEINS ............................................................................................................................................7
TABLE 1.2: ALL SOLENOID PROTEINS AND THEIR PROPERTIES .............................................................................................................17
TABLE 2.1: CONTRIBUTING MEMBER DATABASE INFORMATION FOR INTERPRO RELEASE VERSION 75.0 ..................................................46
TABLE 2.2: THE PERFORMANCE OF NINE DIFFERENT GLOBAL ALIGNMENT METHODS ASSESSED ON 2270 ARMADILLO REPEAT SEQUENCES .................................................................................68
TABLE 2.3: TOP THREE MOST COMMON PHYSICOCHEMICAL PROPERTIES AT EACH ALIGNMENT POSITION FOR THE ARMADILLO REPEAT MSA IN FIGURE 2.9 ..................................................................................................................75
TABLE 2.4: DIFFERENCES IN PRSIIKER BETWEEN START AND MIDDLE TPRSES .................................................................................................80
TABLE 2.5: ODDS RATIO TABLE FOR VARIANTS IN AN MSA ...............................................................................................................................94
TABLE 3.1: THE PERFORMANCE OF EIGHT DIFFERENT GLOBAL ALIGNMENT METHODS ASSESSED ON SEED TPR SEQUENCES .................................................................................................................116
TABLE 3.2: THE NUMBER OF OTHER MATCHES SCORING BELOW THE BIT SCORE THRESHOLD AND THEIR LABELS ......................................................................................................................122
TABLE 3.3: TOP THREE MOST FREQUENT AMINO ACIDS IN EACH ALIGNMENT COLUMN OF THE TPR MSA IN FIGURE 3.5 .........................................................................................................................126
TABLE 3.4: TOP THREE MOST COMMON PHYSICOCHEMICAL PROPERTIES AT EACH ALIGNMENT POSITION FOR THE TPR MSA IN FIGURE 3.5 .............................................................................................................130
TABLE 3.5: DIFFERENCES IN PRSIIKER AND SCALED MES BETWEEN START AND MIDDLE TPRSES ..............................................................................138
TABLE 3.6: RESIDUE PAIRS INVOLVED WITH INTRA-TPR CONTACTS AND SUPPORTING EOCs .............................................................................141
TABLE 3.7: RESIDUE PAIRS INVOLVED WITH INTER-TPR CONTACTS AND SUPPORTING EOCs ...........................................................................142
TABLE 3.8: MEDIAN RSA VALUES FOR ALL THREE REPEAT ARCHITECTURES ...........................................................................................................148
TABLE 3.9: CLINVAR VARIANTS IN TPRS AND THE VARIANT TYPES ......................................................................................................................151
TABLE 4.1: CLASSIFICATIONS OF POSITIONS WITH MEDIAN RSA ..............................................................................................................................171
TABLE 5.1: TOP THREE MOST FREQUENT AMINO ACIDS IN EACH ALIGNMENT COLUMN OF THE ARMADILLO MSA........................................................................................................................................194
TABLE 5.2: RESIDUE PAIRS INVOLVED WITH INTRA-ARMADILLO CONTACTS AND THEIR SUPPORTING EOCs .......................................................206
TABLE 5.3: RESIDUE PAIRS INVOLVED WITH INTER-ARMADILLO CONTACTS AND THEIR SUPPORTING EOCs .......................................................208
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>APC</td>
<td>Anaphase Promoting Complex</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>CES</td>
<td>Contact Enrichment Score</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy Number Variant</td>
</tr>
<tr>
<td>Cryo-EM</td>
<td>Cryogenic Electron Microscopy</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-Separated Values</td>
</tr>
<tr>
<td>dArmRP</td>
<td>Designed Armadillo Repeat Proteins</td>
</tr>
<tr>
<td>DARPins</td>
<td>Designed Ankyrin Repeat Proteins</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>DSSP</td>
<td>Dictionary of Secondary Structure of Proteins</td>
</tr>
<tr>
<td>EoC</td>
<td>Evidence of Contact</td>
</tr>
<tr>
<td>ExAC</td>
<td>Exome Aggregation Consortium</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>gnomAD</td>
<td>Genome Aggregation Database</td>
</tr>
<tr>
<td>GNU</td>
<td>GNU’s Not Unix!</td>
</tr>
<tr>
<td>GO-ESP</td>
<td>&quot;Grand Opportunity&quot; Exome Sequencing Project</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-Wide Association Study</td>
</tr>
<tr>
<td>HAT</td>
<td>Half-A-TPR</td>
</tr>
<tr>
<td>HEAT</td>
<td>Huntingtin, Elongation factor 3, protein phosphatase 2A, TOR1 kinase</td>
</tr>
<tr>
<td>HER2</td>
<td>Human Epidermal growth factor Receptor-2</td>
</tr>
<tr>
<td>HGMD</td>
<td>Human Gene Mutation Database</td>
</tr>
<tr>
<td>HSP</td>
<td>Heat-Shock Protein</td>
</tr>
<tr>
<td>IFIT</td>
<td>Interferon-Induced protein with Tetratricopeptide repeats</td>
</tr>
<tr>
<td>JI</td>
<td>Jaccard Index</td>
</tr>
<tr>
<td>LRR</td>
<td>Leucine-Rich Repeat</td>
</tr>
<tr>
<td>MAb</td>
<td>Monoclonal Antibody</td>
</tr>
<tr>
<td>MCC</td>
<td>Matthews Correlation Coefficient</td>
</tr>
<tr>
<td>MES</td>
<td>Missense Enrichment Score</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger Ribonucleic Acid</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MSA</td>
<td>Multiple Sequence Alignment</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-Generation Sequencing</td>
</tr>
<tr>
<td>NHLBI</td>
<td>National Heart, Lung, and Blood Institute</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>PDA</td>
<td>Protein Domain Architecture</td>
</tr>
<tr>
<td>PDBe</td>
<td>Protein Data Bank in Europe</td>
</tr>
<tr>
<td>PFTB</td>
<td>Prenyltransferase Repeats</td>
</tr>
<tr>
<td>PP5</td>
<td>Protein Phosphatase 5</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein-Protein Interaction</td>
</tr>
<tr>
<td>PPIES</td>
<td>Protein-Protein Interaction Enrichment Score</td>
</tr>
<tr>
<td>PPR</td>
<td>Pentatricopeptide Repeat</td>
</tr>
<tr>
<td>PRA</td>
<td>Protein Repeat Architecture</td>
</tr>
<tr>
<td>PTS1R</td>
<td>Peroxisomal Targeting Signal 1 Receptor</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root-Mean-Square Deviation</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RSA</td>
<td>Residue Solvent Accessibility</td>
</tr>
<tr>
<td>RSCC</td>
<td>Real-Space Correlation Coefficient (RSCC)</td>
</tr>
<tr>
<td>RSR</td>
<td>Real-Space R</td>
</tr>
<tr>
<td>SG</td>
<td>Sequence Group</td>
</tr>
<tr>
<td>SHC1</td>
<td>Sporulation-specific Homolog of CSD4</td>
</tr>
<tr>
<td>SLR</td>
<td>Sel1-like Repeat</td>
</tr>
<tr>
<td>SNP</td>
<td>Single-Nucleotide Polymorphism</td>
</tr>
<tr>
<td>SSE</td>
<td>Secondary Structure Element</td>
</tr>
<tr>
<td>TLR</td>
<td>Toll-Like Receptor</td>
</tr>
<tr>
<td>TPR</td>
<td>Tetratricopeptide Repeat</td>
</tr>
<tr>
<td>TR</td>
<td>Tandem Repeats</td>
</tr>
<tr>
<td>TSV</td>
<td>Tab-Separated Values</td>
</tr>
<tr>
<td>VLR</td>
<td>Variable Lymphocyte Receptors</td>
</tr>
<tr>
<td>WES</td>
<td>Whole-Exome Sequencing</td>
</tr>
<tr>
<td>WGA</td>
<td>Whole-Chromosome Aneuploidy</td>
</tr>
<tr>
<td>WGD</td>
<td>Whole-Genome Duplication</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole-Genome Sequencing</td>
</tr>
<tr>
<td>XLID</td>
<td>X-Linked Intellectual Disability</td>
</tr>
</tbody>
</table>
XML
eXtensible Markup Language
Acknowledgements

I would like to thank Geoff Barton for his unwavering guidance and encouragement for the duration of this project while also showing me invaluable insights into the world of structural biology. Geoff always provided support and patience that cannot be underestimated. I am deeply indebted to Stuart MacGowan for mentoring me early in my project, and I am grateful to him for showing me the ropes. Stuart always supported me and gave me a profound belief in my abilities to become a competent bioinformatician. I would like to express my deepest gratitude to all the students who allowed me to supervise and teach them, particularly Javier S. Utges, Thibault Rosazza, and Noah J. M. Dietrich. I cannot begin to express my thanks to my close friend and colleague, Javier, for sharing the same curiosity in the work we accomplished together. He was instrumental in developing my work, and I am thankful for his company. In addition, I would like to thank everyone else in the Division of Computational Biology, particularly Jim Procter, Benedict Soares, Mungo Carstairs, and Suzanne Duce for the technical support and helpful suggestions with Jalview. For the useful data analysis insight and their witty humour, Nick Schurch, Marek Gierlinski, and James Abbot. In addition, Neil Thomson, Callum Ives, and Dominik Gurvic for the friendship over the years. I would like to express my most profound appreciation to my committee for their support and insight over the years. I want to thank my closest friends, Amanda, Hannah, and Glen, for their unwavering moral support. I am very grateful to my mum, dad, and brother for their love and support. Finally, I would like to thank the Wellcome Trust for funding my PhD and the IT service of the University of Dundee for the support and maintenance of the HPC infrastructure.
STATEMENT

School of Life Sciences, University of Dundee

I certify that Maxim Ivanov Tsenkov has satisfied all the terms and conditions of the relevant Ordinance and Regulations to qualify in submitting this thesis, entitled ‘Exploring human population variation in repeat proteins’, in application for the degree of Doctor of Philosophy.

Date: March 2022

Supervisor: Prof. Geoffrey J. Barton
DECLARATION

School of Life Sciences, University of Dundee

I hereby declare that the work described in this thesis, entitled 'Exploring human population variation in repeat proteins', is my own; that I am the author of this thesis; that it has not previously been put forward in submission for any other degree or qualification; and that I have consulted references herein.

Date: March 2022

Candidate: __________________________________________

Maxim Ivanov Tsenkov
Abstract

Evolutionarily conserved sites in an alignment of protein homologues provide insight into the structural and functional pressures on those proteins. Likewise, global patterns of genetic variation across the human genome can reflect the functional and evolutionary constraints in protein-coding regions. Databases of genetic variation in the human population are essential resources to interpret variants causing rare Mendelian diseases clinically. Disease-causing variants cluster at regions of functional or structural importance in proteins, whereas missense variants are distributed in regions tolerant of amino acid substitutions. However, most human sequence residues in known protein domains have zero variants, so a comparative picture of genetic variation between individual protein residues in the human proteome is not possible.

This situation has been rectified by incorporating human population variant data into conventional sequence analysis, resulting in a larger number of variants that can be used in multiple sequence alignments, thus boosting statistical power. The aim was to understand their relationship at consensus positions of multiple sequence alignments across all protein families. Sites under selective pressure in the human population highlight positions with structural and functional importance that dictate domain evolution, as observed in protein residue conservation.

Unlike globular protein domains with unique folds, other proteins have evolved through periodic sequences of variable arrays duplicated adjacent to one another, termed repeats. Repeat families were focused on to boost the power of this analysis, as repeats present many more copies than any other domain within a single species: Tetratricopeptide Repeat, Ankyrin, and Armadillo repeat families. Repeats are widely distributed across kingdoms and organised in tandem arrays present in numerous proteins involved in many fundamental cellular
processes. These repeats belong to a class of solenoid domains that fold into elongated solenoid structures, where repeats require one another to maintain the overall structure. They mediate protein–protein interactions and assemble multi-protein complexes.

A computational framework, Repeat Analysis Method (RAM), was developed to implement tools and methods in a modularised approach to study repeats. RAM retrieves annotations and analyses the organisation of repeats with protein repeat architecture (PRA). Sequence analysis was carried out, together with the different structural and functional pressures at each position in repeat alignments, as determined by amino acid conservation. The human population variant data from the Genome Aggregation Database (gnomAD) was analysed by aggregating and counting missense variants along each alignment column to identify positions constrained in the human population data. The relationship between residue conservation and the missense variant profile of each repeat alignment was studied. These data were supported by quantitative structural analyses of all available three-dimensional structures from the Protein Data Bank in Europe (PDBe). Positions important for the structural fold of the repeat domains and sites enriched for protein–substrate interactions were identified because they were under constraint in the human population.

This novel cross-disciplinary approach has validated known positions and identified new positions of importance that conventional sequence and structural analysis alone could not. These studies have also expanded the understanding of the key sites in each repeat family, reflecting the general relationship between homologous conservation and population constraint in most protein domains. The analysis demonstrates how the conservation and variation features are constrained by structure and thus can be used to infer structural features.
Chapter 1

Introduction

1.1 Preface

This chapter reviews some of the efforts to understand how we tackle big data and the relationship between protein sequence and structures. Then, the focus narrows towards understanding this relationship further in repeat proteins and understanding how human population variation impacts proteins and repeat.

1.2 Background

Since discovering the structure of DNA (Watson and Crick, 1953), their work has established the central dogma of molecular biology we know today – the coding sequence of DNA is transcribed to RNA which translates to proteins (Crick, 1970). Thirty-seven years later, their work has set in motion the biggest undertaking in human history – the human genome project (HGP) (Collins et al., 2003). An inward voyage to understand how genomic DNA defines species and us as individuals. Following the completion of the HGP, patterns of human genetic diversity raised implications for studying the human evolutionary history and disease (Tishkoff and Verrelli, 2003). The last two decades have seen enormous efforts in sequencing hundreds of thousands of individuals (Lek et al., 2016, Consortium, 2010). These endeavours have helped us understand the underlying genetic variation that has shaped human genomes,
which is the culmination of the entire human evolution (Wohns et al., 2021). With the dramatic growth of genomic data, many in the community have tried to make sense of millions of protein sequences (Consortium, 2017).

Proteins form the basis of living cells where their primary function is to transfer and process information in nearly every cellular process (Bray, 1995, Levitt, 2009). Proteins extensively function through protein-protein interactions (PPIs) (Nooren and Thornton, 2003, Russell et al., 2004, De Las Rivas and Fontanillo, 2010). Proteins make up the main machinery, known as the proteome, which reflects the dynamic state and underlying biological processes of a cell, tissue, and organism (). Natural selection through genetic variation and sequence variation within and between species introduces the variability that drives and shapes the evolution of all living organisms (Pool et al., 2010, Simonti and Capra, 2015). Understanding the influence of random mutations on protein folding and stability (Lorch et al., 2000, Koukouritaki et al., 2007), protein function (Yamada et al., 2006), and PPIs (Yates and Sternberg, 2013, David et al., 2012) is critical for predicting their functional impact (Ramensky et al., 2002, Ng and Henikoff, 2006, Reva et al., 2011, Sim et al., 2012). The consequences of genetic variation on the function and evolution of proteins are explored here in this thesis.

1.3 The Era of Big Data

The life sciences and biomedical domain are entering a new era of protein science. Next-generation sequencing (NGS) has revolutionised how DNA information is accessed in ways that were previously not possible (Goodwin et al., 2016). Consequently, the emergence of new and varied technologies have paved the way for new applications, such as real-time pathogen DNA monitoring (Gardy and Loman, 2018), routine clinical DNA sequencing (Van El et al., 2013), and large-scale population studies (Consortium, 2010). The ability to routinely sequence entire genomes (Köser et al., 2012) or determine millions of environmentally derived sequences (Yilmaz and Singh, 2012, Quince et al., 2017) help us peer into numerous and increasingly diverse organisms for new protein sequences.
Great strides have been also made in the experimental techniques in structural biology. Cryogenic electron microscopy (cryo-EM) single-particle analysis of large macromolecular assemblies has revolutionised structural biology, achieving near-atomic resolution of any size and rivalling the gold-standard X-ray crystallography method (Zhou, 2008, Bai et al., 2015).

The research community has focused on developing and refining technologies that increase the depth and quality of raw data. In contrast, others have employed computational approaches to gain new insights from existing data (Keskin et al., 2008), developing the ability to functionally annotate new sequences (Finn et al., 2017) automatically with the help of sequence profiles (Eddy, 2011) from multiple sequence alignments (MSA), and predict their function (Lee et al., 2007). Others have developed computational methods that are trained on solved structures to predict secondary structure elements (SSEs) of novel sequences (Drozdetskiy et al., 2015). But more recently, atomic resolution of three-dimensional structures has been achieved with high accuracy, even when no homologous structures are available (Ovchinnikov et al., 2018, Jumper et al., 2021).

However, these massive strides are mired in a deluge of data that we cannot easily cope with. Bioinformatics is an interdisciplinary field that employs mathematical, statistical, and computational methods to analyse the vast quantities of biological data from genomics and proteomics studies. The information gap between experimental data and biological knowledge grows exponentially, all the while trying to organise it (Consortium, 2017). The advent of NGS has generated a surge in discovery of novel proteins, and the scientific community cannot experimentally characterise them at the same pace (Consortium, 2017). Moreover, advances in high-throughput technologies generate new data types in genomics (Mardis, 2013, Liu et al., 2012, Goodwin et al., 2016) and proteomics (Tyers and Mann, 2003, Cho, 2007) with increasing complexity and depth. The challenge here is to capture the increasing breadth of information to be full exploited and interpretable by the community. At the same time, keeping
up with progress in experimental techniques that continue to provide increasing breadth and depth of information on the structure and function of proteins.

1.3.1 Functional annotation of protein sequences

The universal protein knowledgebase (UniProtKB) is a long-standing collection of databases that catalogue over half a million sequences (Consortium, 2017). UniProt is a comprehensive resource that provide many protein sequences and associated annotation data. UniProtKB consists of two sections, Swiss-Prot and TrEMBL. Protein entries in UniProtKB/Swiss-Prot, which make up the reviewed section of UniProtKB, have been created and are maintained by an expert bio-curation team, who critically review them with experimental and predicted data. The remainder of entries are automatically annotated in UniProtKB/TrEMBL in response to the increasingly large sequencing data from genome projects. These protein entries are computationally analysed with automatic annotation and classifications, which are kept separately from Swiss-Prot.

1.3.2 Predictive diagnostic models

The development of automated sequence analysis methods allows the transfer of information from few experimentally characterised sequences to uncharacterised homologue sequences, known as protein signatures (Blum et al., 2021). These methods include hidden Markov models (HMMs) (Eddy, 1995), patterns, profiles, and fingerprints. Several databases have specialised in employing some of these methods in Figure 1.1: Pfam (El-Gebali et al., 2019), SMART (Letunic and Bork, 2018), PROSITE Patterns and Profiles (Sigrist et al., 2012), PRINTS (Attwood et al., 2012), CATH-Gene3D (Sillitoe et al., 2019), the Conserved Domains Database (CDD) (Lu et al., 2020), HAMAP (Pedruzzi et al., 2015), PANTHER (Mi et al., 2019), PIRSF (Nikolskaya et al., 2006), Structure-Function Linkage Database (SFLD) (Akiva et al., 2014), SUPERFAMILY (Pandurangan et al., 2019) and TIGRFAMs (Haft et al., 2012).
1. Introduction

1.3 The Era of Big Data

Figure 1.1: Schematic representation of the member databases making up InterPro. InterPro integrates thirteen member databases, which are categorised based on the signature method they use and Biological Entity they annotate. Courtesy from the InterPro webpage [https://www.ebi.ac.uk/interpro/about/interpro/] (Blum et al., 2021).

TrEMBL uses a resource called InterPro that permits searches of several libraries simultaneously and functionally annotate regions of unreviewed sequences (~230 million) automatically (Consortium, 2015). The InterPro database is a meta-resource that combines a non-redundant set of protein signatures from 14 different domain and motif databases integrated into a single searchable resource, classifies proteins into families and identifies functional regions of proteins (Blum et al., 2021). InterPro is the largest source of automatic annotation of sequences in UniProtKB (Consortium, 2018). The released version InterPro 75.0 is summarised in Table 1.1 and broken down further later in Table 2.1, showing the current statistics and numbers for this release. InterPro follows a hierarchical protein classification system, where one or more member database signatures representing the same protein family,
domain, or site is assembled by InterPro into an Entry. InterPro classifies these Entries into different Biological Entities: Homologous Superfamilies, represent a large and diverse family with shared three-dimensional structures; Families are a group of proteins with a common evolutionary origin, which is reflected in similar functions and shared sequence and structure homology; Domains are distinct functional and structural units that contributes to the role of a protein; Repeats are arrays of short repeating stretches of amino acid sequences with regular secondary structures; Sites are groups of amino acids that confer specific characteristics upon a protein, such as an active site or a post-translational modification site.
1.3 The Era of Big Data

Table 1.1: Contents and coverage of the InterPro protein matches for all UniProtKB and UniParc proteins. InterPro release version 75.0 contains 36,872 entries based on the InterPro Entries. Sourced from the release notes of InterPro [https://www.ebi.ac.uk/interpro/release_notes/].

<table>
<thead>
<tr>
<th>Biological Entity</th>
<th>Number of protein matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homologous Superfamily</td>
<td>3152</td>
</tr>
<tr>
<td>Family</td>
<td>21886</td>
</tr>
<tr>
<td>Domain</td>
<td>10604</td>
</tr>
<tr>
<td>Repeat</td>
<td>317</td>
</tr>
<tr>
<td>Site</td>
<td></td>
</tr>
<tr>
<td>Active Site</td>
<td>132</td>
</tr>
<tr>
<td>Binding Site</td>
<td>76</td>
</tr>
<tr>
<td>Conserved Site</td>
<td>688</td>
</tr>
<tr>
<td>PTM</td>
<td>17</td>
</tr>
</tbody>
</table>

1.3.3 Multiple sequence alignments
A multiple sequence alignment (MSA) refers to the process and result of the alignment of homologous sequences, such as RNA, DNA, or proteins. The evolutionary history of a protein, and the functional and structural relationship amongst homologous sequences can be inferred from an optimally constructed MSA (Wallace et al., 2005). Until the mid 1980s, MSAs were done by hand, because the traditional dynamic programming algorithm proved to be too slow when there were more than three sequences being aligned (Needleman and Wunsch, 1970). Since then, the last 40 years have seen the development of many alignment methods employing different kinds of heuristics to replicate evolution (Chatzou et al., 2016). The most popular approach uses progressive alignment construction, also known as a hierarchical or tree method, originally developed by Feng and Doolittle, 1987. Iterative methods work similarly to progressive methods by realigning initial and new sequence of the growing MSA a certain number of times using a guide tree or random splits, such as the MUSCLE alignment method (Edgar, 2004). Hidden Markov Modelling (Eddy, 1995) is a probabilistic model that assigns the likelihood of amino acid and gaps to statistically summarise an MSA, known as an HMM.
profile. Other approaches include consensus-based methods, that build an optimal MSA from multiple MSAs built with different alignment methods (Collingridge and Kelly, 2012).

MSAs are essential precursors for downstream bioinformatics applications, such as phylogenetic tree constructions (Phillips et al., 2000), domain analysis, Hidden Markov Model (HMM) profiles, protein structure and function prediction, and many other applications that have been extensively reviewed over the past two decades (Edgar and Batzoglou, 2006, Kemena and Notredame, 2009, Thompson et al., 2011, Chatzou et al., 2016).

Alignment methods entail a series of algorithmic processes that optimally align a given set of sequences while taking into consideration evolutionary events, such as point mutations, insertions and deletions (indels). Popular alignment methods include T-Coffee (Notredame et al., 2000); ClustalΩ (Sievers et al.); Clustal W (Larkin et al., 2007); MUSCLE (Edgar, 2004); MAFFT (Rozewicki et al., 2019); ProbCons (Do et al., 2005). This is a small selection from nearly a hundred alternative alignment methods developed in the last three decades (Kemena and Notredame, 2009). Other alignment methods, such as AMPS (Barton, 1994) and 3D-Coffee (O’Sullivan et al., 2004) can integrate structural information with normal sequencing data to improve alignments (Wallace et al., 2005).

MSA methods will be challenged in the current genomic era, where they need to handle large sequences with no homologous sequences or available structures. This is especially true, to be able to automatically detect subfamilies within genomic sequences (Abhiman et al., 2006). Methods like MUSCLE (Edgar, 2004), Mafft (Rozewicki et al., 2019), and Kalign (Lassmann and Sonnhammer, 2005b) are currently capable of large-scale alignments (Kemena and Notredame, 2009).

Amino acid variability in an informative MSA encapsulates the evolutionary pressure and mutations observed (Gil and Fiser, 2018). Inferring amino acids patterns can reflect the importance of positions in maintaining the structure and function of a protein fold, which in
turn constrains the amino acid types and their properties. Measuring this property relies on methods scoring residue conservation, where 17 methods were extensively reviewed and compared by Valdar (2002). Taylor (1986) defined a minimal set of physicochemical properties that represent amino acids, which was then developed as a method for scoring residue conservation (Zvelebil et al., 1987). Most scoring methods can highlight the overall similarity at each position in an alignment but overlook the similarities of subfamilies within diverse protein families. The AMAS program (Analysis of Multiply Aligned Sequences) quantitatively measures differences and similarities between subfamilies (Livingstone and Barton, 1993).

It is important to acknowledge the differences between biological concepts of informativeness (or biological correctness) and accuracy when it comes to assessing an MSA (Chatzou et al., 2016, Gil and Fiser, 2018). One of the key biological issues surrounding MSAs is the ability to achieve informativeness, where equivalent positions with a functional or structural role can be inferred from an informative MSA. As for accuracy, it entails minimising alignment errors (Gil and Fiser, 2018). Assessing the quality of an MSA has often been down to benchmarking the accuracy of alignment methods. For example, measuring the quality of an MSA with a conservation score (Lin et al., 2011), consensus agreement across multiple alignment methods (Lassmann and Sonnhammer, 2005a), or more traditionally, comparing against gold standard reference alignment databases derived from structures, such as BAliBASE (Thompson et al., 2005).

However, current methods lack the diversity to accommodate for biologically informative MSAs (Gil and Fiser, 2018). This underlying issue is a necessary step towards understanding a term coined in late 80s – ‘twilight zone’. This term is used to describe a phenomenon where homologues sequences with low sequence identify can still fold into the same and conserved structural fold (Chothia and Lesk, 1986, Rost, 1999, Flores et al., 1993).
This tends to be in large families with high sequence diversity, where the sequences have fewer than normal conserved positions whilst preserving a largely similar overall structure. Proteins with low sequence identities tend to have conserved structures, even as low as 15% residue identity, that have similar folds and functions (Orengo et al., 1993, Orengo et al., 1997). Examples include immunoglobulins, crucial for the function of the immune system (Chothia et al., 1989), and SH2 and SH3 domains that aid in the signal transduction of receptor tyrosine kinase pathways (Russell and Barton, 1992, Livingstone and Barton, 1993).

1.3.4 The Protein Data Bank

Similar to UniProt, structures solved for protein sequences also need to be organised for researchers to access. The Protein Data Bank (PDB) is a major structural repository for structural data that catalogues, processes, and archives structures solved by various methods (Berman et al., 2003, Berman et al., 2007, Velankar et al., 2010). Since its inception, it has considerably increased the number of structural data deposited. Currently, it is exponentially increasing, where the PDB currently archives approximately 180,000 structures. Furthermore, the Electron Microscopy Data Bank (EMDB) (Lawson et al., 2016) stores around 19,000 structures solved by either electron crystallography, single-particle analysis, and electron tomography (Patwardhan, 2017).

The most popular experimental techniques for three-dimensional structure determination are X-ray crystallography (Smyth and Martin, 2000, Wlodawer et al., 2008), Nuclear Magnetic Resonance (NMR) (Wüthrich, 1989, Luca et al., 2003, Kaplan et al., 2016), and cryo-EM (Bartesaghi and Subramaniam, 2009, Bonomi et al., 2018). X-ray crystallography methods can gain atomic resolution (1.5 Å or lower), whereas cryo-EM techniques have poorer resolutions of more than 3Å. However, the cryo-EM method has the advantage of solving large macromolecular complexes that cannot be achieved with X-ray crystallography. Thus, many studies have often solved large protein complexes that employ both methods, where often there
are already X-ray solved structures archived in the PDB to determine cryo-EM structures with high resolution (Trabuco et al., 2009, Chen and He, 2018).

1.3.5 Structure classification
Domains are structural, functional, and evolutionary building blocks that represent the most basic unit of a structural fold (Light and Elofsson, 2013, Ponting and Russell, 2002). Furthermore, domains are self-contained and spatially distinct that can fold and function in isolation (Ponting and Russell, 2002). Domains are evolutionary modules, where new sequences tend to be adapted from pre-existing sequences with common structural motifs (Jacob, 1977). In contrast, some structural folds have been reinvented by convergent evolution, giving rise to similar local structures (Lupas et al., 2001). Global similarities between domains on a sequence or structural level may suggest an evolutionary relationship.

Class, Architecture, Topological, and Homologous Superfamily (CATH) (Orengo et al., 2002) and Structural Classification of Proteins (SCOP) (Lo Conte et al., 2000) are two of the main hierarchical domain classification systems. The structural hierarchy of CATH has four levels. Class refers to the general composition of secondary structure elements (SSEs), such as mainly α-helices, mainly β, or α-β mix. Each Class is further classified to Architecture, which represents the general arrangement of SSEs in three-dimensional space, regardless of the order they are linked in. Topology describes the domains that have share a similar number of SSEs and the order in which they are linked. Lastly, Homologous Superfamily groups structures with highly similar structures and functional similarity.

1.3.6 Protein domain architecture
Protein domain architecture (PDA) describes the sequential organisation of domains at the primary sequence level of proteins (Lesk, 2001). This information is considered the fundamental level for protein functional complexity. PDA provides an insight into the evolution of proteins that give rise to complex entities (Holm and Sander, 1994).
The understanding of the evolution and organisation of PDA has been extensively studied (Björklund et al., 2005, Levitt, 2009, Hsu et al., 2013). Although there are single-domain proteins, they tend to be found in combination with other domains (Ponting and Russell, 2002). The prevalence of multi-domain proteins rises with increasing complexity of species, where eukaryotes have more than bacteria (Ekman et al., 2005). Proteins can gain new functions and specificities through domain duplication, insertion, deletion, recombination, and exchange (Forslund and Sonnhammer, 2012).

1.3.7 Methods for structure comparison

There are many different sequence structural elements (SSEs), such as α-helix, π-helix, 3-helix, β-sheets, etc. (Kabsch and Sander, 1983). To understand how SSEs come together to fold correctly, there are two main methods of protein structure comparison: distance-based and contact-based.

Distance-based methods are dependent on the distance measurements between reference points in a structure model for a superimposition (Kufareva and Abagyan, 2011). While distance-based methods use three-dimensional models, an alternative to this simplifies it with two-dimensional representations known as contact matrices. A contact matrix is a binary, and more reduced, representation of all residue pair contacts in a structure (Tradigo, 2013). The community has frequently trained on contact matrices to model predicted tertiary structure of proteins (Abu-Doleh et al., 2012, Hou et al., 2019).

1.3.8 The most common classes of protein architectures

Domains are evolutionary modules, where new sequences tend to be adapted from pre-existing sequences (Jacob, 1977). Protein domains are a combination of structural motifs that often share general combinations and arrangements of SSEs with other proteins on the Architecture level (Orengo et al., 1997). For example, α–β sandwiches have very similar folds and are capable of adapting active sites to accommodate different functional requirements (Orengo et al., 1993). Globular domains are usually aperiodic and tend to combine existing or
favourable structural folds; reoccurring and favourable folds have evolved differently, with a repeating structural motif.

The earliest suspicions concerning the profusion of repeats were first raised by Gerstein (1997). Since then, tandem-repeat domains have been confirmed as one of the most common classes of protein architectures (reviewed by Heringa, 1998). They are found throughout subcellular locations in at least 14% of all proteins and are ubiquitous across the tree of life (Marcotte et al., 1999, Kamel et al., 2021). With growing genomic data available, Kajava (2012) laid out and classed all types of protein repeats that range from single amino acids to domains of 60-or-more residues.

Similar to the hierarchical class of CATH, Kajava (2012) put forward a classification system for repeat families. This was further developed by Di Domenico et al. (2014) for RepeatsDB. The RepeatsDB is a database of classifications and annotations for tandem repeated protein structures from the PDBe (Paladin et al., 2021). There are five classes: Class I are crystalline aggregates of unlimited size; Class II are fibrous proteins, such as collagens (Rich and Crick, 1955, Bella et al., 1994); Class III, elongated structures that form linear solenoid domains or closed toroid folds, such as Armadillo (Andrade et al., 2001) and WD40 repeats (Smith et al., 1999); Class IV, closed structures that that need one another to fold correctly, such as TIM-barrels (Triose-Phosphate Isomerase) (Alvarez et al., 1998); Class V are repeats independent of one another and capable of folding separately linked together that are described as “beads on a string”, such as Zn-finger domains (Lee et al., 1989).

Repeats belonging to Class III form elongated structures in which repeats require one another to maintain structure (Kajava, 2012, Kobe and Kajava, 2000). The most commonly studied repeat families are Tetratricopeptide Repeats (TPRs) (D’Andrea and Regan, 2003), Ankyrins (Voronin and Kiseleva, 2008), Leucine-Rich Repeats (LRR) (Enkhbayar et al., 2004), Armadillos and Huntingtin, elongation factor 3 (EF3), protein phosphatase 2A (PP2A),

13
and the yeast kinase TOR1 (HEAT) (Andrade et al., 2001). TPRs, Ankyrins, and Armadillos are observed across a wide range of species in Eukaryotes, Prokaryotes, and Archaea (Jernigan and Bordenstein, 2015, Kamel et al., 2021).

Unlike complex globular folds of domains that occur stably in isolation (Ponting and Russell, 2002), repeats are functionally versatile structural units that fold collectively and reside adjacent to one another in the primary sequence; termed tandem repeats. They individually adopt simple and regular arrangements of SSEs. But together, they form three-dimensional assemblies of linear (open) and toroid (closed) superstructural conformations that cannot fold in isolation without the support of adjacent repeat (Kajava, 2012, Kobe and Kajava, 2000).

Repeats are considered an economic way of evolving complex structures. Adding or removing a repeat will make little to no difference, as they can retain a stable and functional fold (Kobe and Kajava, 2000, Kajava, 2012). The relative ease of evolution and profusion throughout the proteome reflects their abundance across species (Jernigan and Bordenstein, 2015, Kamel et al., 2021). The mechanisms that shape repeats include exon duplication and rearrangement. In plants, repeats were observed to reside within individual exons (Schaper and Anisimova, 2015). Therefore it is likely that the region with adjacent repeats was established due to exon shuffling (Bjorklund et al., 2006) and not replication slippage commonly found with TRs in genomic sequences (Levinson and Gutman, 1987, Ellegren, 2000, Mirkin, 2006).

Paladin et al. (2020) demonstrated, on proteins from the RepeatsDB, how some repeat families had exon boundaries residing between repeats. Not all proteins displayed this, and the repeats in structure appeared disjointed from exons. It is likely that some repeat proteins did not originally evolve in the species in which it is found. Instead, there has likely been an ancestral repeat expansion before speciation occurred (Schaper et al., 2014). Thus, it may have lost its original exon junctions that brought about the organisation of the repeats in certain
proteins. While some efforts to understand how repeat organisation evolved, others have attempted to quantitatively capture the current repeat architecture in proteins.

The community studying repeats has attempted to visualise and describe the organisation of repeats across most families. In RepeatsDB ‘Browse’ page for a repeat family, this database visualises the query with a summary page that visualises a bar plot summarising the number of repeats in a region. A bar plot for a repeat family visualises the number of regions (number of domains with repeats) and the number of units (repeats) that make up each domain.

For the TPR family, D'Andrea and Regan (2003) showed the distribution of TPR domain lengths across multiple species. This proteome-wide analysis captures the total number of TPRs to understand how the number and distribution of TPRs contrasts between species. Like the TPR analysis, Mosavi et al. (2004) also explored the total number of Ankyrins proteome-wide.

Although not explicitly mentioned, some work inadvertently discusses the repeat architecture of certain repeat families. Such work tends to revolve around reviews that carry out a qualitative proteome-wide analysis of the function and evolution of a repeat family. For example, Coates (2003) listed all Armadillo-containing proteins known at the time of writing. This work discusses all the proteins shared across kingdoms and their functions. Paired with them are the PDA of domains and the repeat architecture of Armadillo repeats. A further example is the review by Tewari et al. (2010) that assessed the Armadillo repeat architecture within well-established model organisms. Additionally, Tewari et al. (2010) made a case for studying novel Armadillo-containing proteins in under-studied organisms to learn more about the diversity of this repeat family and the arrangement of Armadillo repeats in them.

Individual protein-specific studies have even used the term ‘repeat architecture’ when crudely referring to any element of repeat organisation. Knutson (2010) reannotated TPR and HEAT repeats in serine/threonine-protein kinase TOR (Target of Rapamycin) protein
Introduction

1.4 Repeat families

1.4.1 Solenoid proteins

Solenoid proteins are so named because the polypeptide chain loops on itself to form a solenoid. In fact, depending on the how adjacent repeats stack, the solenoid structure can continue looping by forming a superhelix. While solenoids are the appropriate name, they have been previously described as coils, spirals, superhelices, coiled folding domains, amongst others (Reviewed in Kobe and Kajava, 2000). The structural geometry of repeat proteins, and how repeats are categorised according to their structural properties and sequence, will be reviewed here.

1.4.2 Topological arrangements of helical repeat families

Repeat proteins can be characterised in terms of their structural fold. Repeats can be defined by specific arrangements and combinations of SSEs, termed structural motifs. The topological arrangements of the SSEs defines each repeat family. The simplest solenoid proteins have basic structural motifs comprising two SSEs with either α/α, β/β, or α/β units. More complex organisations include three or even four SSEs in a unit.
Table 1.2: All solenoid proteins and their properties. Table adapted from Kobe and Kajava (2000).

<table>
<thead>
<tr>
<th>Structural Unit</th>
<th>Sequence Repeat</th>
<th>Protein Fold</th>
<th>Twist</th>
<th>Curvature</th>
<th>Representative Protein</th>
<th>PDB Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>α/α</td>
<td>Weak</td>
<td>Right-handed α-2-solenoid</td>
<td>Irregular, mostly small right-handed</td>
<td>Large</td>
<td>Lipovitellin-phosvitin (lamprey yolk)</td>
<td>-</td>
</tr>
<tr>
<td>α/α</td>
<td>Weak</td>
<td>Right-handed α-2-solenoid</td>
<td>Irregular, mostly right-handed</td>
<td>Large, irregular</td>
<td>Soluble lytic transglycosylase (Escherichia coli)</td>
<td>1QSA</td>
</tr>
<tr>
<td>α/α</td>
<td>Weak</td>
<td>Right-handed α-2-solenoid</td>
<td>Very small</td>
<td>Small</td>
<td>Clathrin heavy chain, residues 1210–1516</td>
<td>1B89</td>
</tr>
<tr>
<td>α/α</td>
<td>Weak</td>
<td>Right-handed α-2-solenoid</td>
<td>Irregular, mostly right-handed</td>
<td>Moderate</td>
<td>Clathrin heavy chain, N-terminal fragment</td>
<td>1BP0</td>
</tr>
<tr>
<td>α/α</td>
<td>Weak</td>
<td>Right-handed α-2-solenoid</td>
<td>Right-handed</td>
<td>Small</td>
<td>Sec17</td>
<td>1QQE</td>
</tr>
<tr>
<td>α/α</td>
<td>Detectable</td>
<td>Right-handed α-2-solenoid</td>
<td>Mostly left-handed</td>
<td>Small</td>
<td>Protein farnesyltransferase, α-subunit</td>
<td>1FT1</td>
</tr>
<tr>
<td>α/α</td>
<td>Tetratricopeptide repeat</td>
<td>Right-handed α-2-solenoid</td>
<td>Right-handed</td>
<td>Small</td>
<td>Protein phosphatase 5</td>
<td>1A17</td>
</tr>
<tr>
<td>α/α</td>
<td>HEAT repeat</td>
<td>Right-handed α-2-solenoid</td>
<td>Variable</td>
<td>Large</td>
<td>Protein phosphatase 2A PR65/A subunit</td>
<td>1B3U</td>
</tr>
<tr>
<td>α/α</td>
<td>Armadillo repeat</td>
<td>Right-handed α-3-solenoid</td>
<td>Right-handed</td>
<td>Small</td>
<td>β-Catenin</td>
<td>3BCT</td>
</tr>
<tr>
<td>α/3+</td>
<td>Leucine-Rich repeat variant</td>
<td>Right-handed α-3-solenoid (horseshoe)</td>
<td>Very small</td>
<td>Moderate</td>
<td>Leucine-rich repeat variant protein (LRV) (Azotobacter vinelandii)</td>
<td>1LRV</td>
</tr>
<tr>
<td>β/β/α/α</td>
<td>Ankyrin repeat</td>
<td>Right-handed α-3-solenoid (bundle)</td>
<td>Small left-handed</td>
<td>Small</td>
<td>p53-binding protein 53BP2</td>
<td>1YCS</td>
</tr>
<tr>
<td>α/β</td>
<td>Leucine-Rich repeat (R-like)</td>
<td>Right-handed α-β-solenoid</td>
<td>Very small</td>
<td>Large</td>
<td>Ribonuclease inhibitor</td>
<td>1A4Y</td>
</tr>
<tr>
<td>α/β</td>
<td>Leucine-Rich repeat (Typical)</td>
<td>Right-handed α-β-solenoid</td>
<td>Small</td>
<td>Moderate</td>
<td>U2A’ protein from spliceosomal snRNP</td>
<td>1A9N</td>
</tr>
<tr>
<td>3+β</td>
<td>Leucine-Rich repeat (internalin-like)</td>
<td>Right-handed 3-β-solenoid</td>
<td>Small</td>
<td>Moderate</td>
<td>Internalin B (Listeria monocytogenes)</td>
<td>1D0B</td>
</tr>
</tbody>
</table>

Structural Unit: SSE of the repeat. Sequence Repeat: Repeat family (if known) and how degenerate the underlying consensus sequence is. Protein Fold: The handedness of the solenoid protein. Twist: is defined in Section 1.4.3. Curvature: is defined in Section 1.4.3. Representative Protein: Example protein that represents this family. PDB Code: The PDB code of the structure for the corresponding family in that row.
1.4.3 Structural features of solenoid proteins

Contiguous repeats stacked together can form many different superstructures. There can be substantial differences in terms of how adjacent repeats arrange themselves. Repeat organisation involves the stacking interactions of hydrogen-bonding and hydrophobic packing. Furthermore, the sidechain sizes also have an effect; either large hydrophobic and aromatic sidechains, or something as tiny as glycine or alanine.

Kobe and Kajava (2000), put forward three simple structural elements of solenoid proteins that can be used to characterise them: handedness, curvature, and twist.
Table 1.2 shows how these three descriptors define the structural geometry of solenoid proteins from different repeat families. The foremost basic property that characterises solenoid proteins is their handedness. This refers to the direction that the polypeptide chain winds around a superhelical axis of the solenoid protein. The TPR, Ankyrin, and Armadillo proteins are similar in that they all fold into right-handed solenoids.

A further property of solenoid proteins is their twist. The twist of a solenoid is determined by how adjacent repeats pack with one another. Kobe and Kajava (2000) measured the direction of the twist by identifying analogous points of a repeat and tracing these points. This traced line forms a type of helix and, analogous to measuring the handedness, is used to define the direction in which the solenoid protein twists. Both TPR and Armadillo domains have a right-handed twist, whereas Ankyrin domains have a small and left-handed twist (}
The last geometric property is the curvature of solenoid proteins. There is much complexity when defining this property, especially with all helical solenoid proteins that have many repeats. This arises due to differences in the packing properties of adjacent repeats. Amino acids that deviate from the consensus sequence, with different sidechain properties, bring about changes in the curvature of solenoid proteins.

What may shape the curvature of solenoid proteins could be a combination of insertions and deletions (indels), and the expansion and loss of repeats (Paladin et al., 2020). Since repeats are structurally malleable and can tolerate a wide range of amino acids (Figure 1.2), even at conserved sites, they likely evolve around their substrates. Aside from the surface residues that are specialised to recognise substrates, solenoid proteins may change their geometrical properties to accommodate certain substrates. For example, the fourth TPR of a TPR domain in peroxisomal targeting signal 1 receptor, with seven TPRs, helps the TPR domain to fold on itself and form a clamp to bind to.

1.4.4 Protein–protein interactions

Protein interactions form the basis of all protein functions and are essential for living cells. PPIs are fine-tuned interactions orchestrated by specific residues, which give rise to specific protein complexes that are make up a network of interactions (Jones and Thornton, 1996, Nooren and Thornton, 2003, Russell et al., 2004, Pazos and Valencia, 2008, Lehne and Schlitt, 2009). Many, but not all repeat families are specialised in binding to substrates, often employed by multi-domain proteins to bind to a variety of proteins (Zeytuni and Zarivach, 2012, Voronin and Kiseleva, 2008, Perez-Riba and Itzhaki, 2019, Magliery and Regan, 2005). This is discussed in the individual introduction sections of the TPR (Section 3.2), Ankyrin (Section 4.2), and Armadillo (Section 5.3) chapters. Utilising PPIs for use in industry and therapeutic research has been of interest in the research community (Zinzalla and Thurston, 2009, Boersma and Pluckthun, 2011).
The basis of protein engineering is to adjust the biophysical and functional properties of proteins using alternative amino acids and indels. The main goal of modifying proteins in this manner is to fine-tune or modify functions for binding specific targets in industry, research, and therapeutics (Binz et al., 2005, Forrer et al., 2004). The most straightforward approach to protein engineering involves building a statistical summary of each position in an MSA of a protein domain. Consensus-designed proteins are based on selecting frequently observed amino acids from a sequence logo. This method is not fool-proof and is often supplemented with structural and computational analysis (Forrer et al., 2004). Furthermore, the consensus-designed approach contrasts directed evolution (Packer and Liu, 2015) and screening different scaffold candidates with computational modelling (Damborsky and Brezovsky, 2014).

Consensus-design has been successfully implemented in several protein families to achieve higher structural stability, that maintains biological function compared with naturally occurring homologues (Boersma and Pluckthun, 2011, Sternke et al., 2019). This has been demonstrated by increasing the stability of antibodies (Ohage and Steipe, 1999, Knappik et al., 2000, Demarest et al., 2004), SH3 domains (Rath and Davidson, 2000), DNA-binding proteins (Nikolova et al., 1998), Ankyrin-containing proteins (Mosavi et al., 2002), and many other protein families, over the last two decades (Boersma and Pluckthun, 2011, Porebski and Buckle, 2016).

1.4.4.1 Antibodies

Monoclonal antibodies (MAbs) are frequently-used binders that have dominated biotechnology and therapeutics over the last 40 years (Kennedy et al., 2018, Scott et al., 2012). However, MAbs have limitations: they are still produced by traditional immunisation methods that have largely remained unchanged. Furthermore, they are difficult and expensive to manufacture. MAbs also remain a challenge to synthesise as effective drugs due to the lack of high specificity (reviewed in Starr and Tessier, 2019), where every target is a new challenge, and the performance and specificity of each MAb has to be assessed individually from antibody
libraries (Knappik et al., 2000). The last two decades have seen a lot of interest in engineering natural proteins to provide alternatives to traditional antibodies and other small binding proteins (Boersma and Pluckthun, 2011, Binz et al., 2005).

1.4.4.2 Designed repeat proteins

Not all adaptive immune systems mount responses with immunoglobulins: jawless vertebrates possess variable lymphocyte receptors (VLRs) that are LRR-containing proteins (Boehm et al., 2012). Additionally, most plant and animal immune receptors of the innate immune system use LRR-containing proteins (Reviewed in Padmanabhan et al., 2009).

Families of proteins have evolved with repeating sequences that employ a modular strategy. For example, Zinc-finger (Znf) domains function by binding to DNA in a sequence-specific fashion (Jantz et al., 2004), and other classes of Znf domains are capable of binding to RNA (Hall, 2005) and lipid (Gamsjaeger et al., 2007) substrates. Functional DNA-binding proteins based on Znf domains have been designed with varying array lengths to recognise different ranges of DNA sequences. Other, more specific repeat families, include the Transcription Activator-Like Effector (TALE) repeats, in Figure 1.3, that have evolved to bind to DNA elements (Deng et al., 2012, Mak et al., 2012). TALE repeats have also been engineered to bind to select stretches of DNA sequences (Boch et al., 2009, Christian et al., 2010, Zhang et al., 2011).

Scaffolds designed from other naturally occurring repeats may be a novel class of binding proteins capable of binding to any substrate type (Boersma and Pluckthun, 2011, Binz et al., 2005). The hydrophobic core is crucial for the biophysical stability of a repeat-containing protein, whereas its solvent-accessible residues on the surface can bind to a diverse repertoire of substrates. This arrangement permits the two properties of repeat proteins to change independently of each other, thus making them particularly attractive for designing repeat scaffolds.
Successful consensus-designed repeat scaffolds have been solved for multiple repeat families: Consensus-TPR 2 and 3 (CTPR2/3) proteins with TPRs (Main et al., 2003), which were designed with a solvating helix at the C-terminus to assist folding; 3ANK and 4ANK for Ankyrin repeats (Mosavi et al., 2002), with three and four repeats, respectively, that had no additional N- or C-terminus capping repeat to assist in folding and stability; N3C, N4C, N5C, N6C in LRR (Stumpp et al., 2003) that has capping repeats (N and C) with a set number of LRR pairs (3–6), developed from libraries of the mammalian ribonuclease inhibitor; The \( Y^{\text{II}} M^{\text{II}} A^{\text{II}}, Y^{\text{III}} M^{\text{III}} A^{\text{III}}, \) and \( Y^{\text{III}} M^{\text{III}} A^{\text{III}} \) in the Armadillo repeat family (Parmeggiani et al., 2008, Madhurantakam et al., 2012) that follows the naming system \( Y Z M X A Z \), the \( X \) signifying the number of mutated Armadillo repeats (M) obtained by computational means; \( Y \) and \( A \) denote the N-terminal capping repeat from the yeast importin-\( \alpha \) and artificial and C-terminus capping repeat, which are numbered based on the Z generation capping repeat, designed with molecular dynamic simulations.

There have been clear successes with designed repeat proteins, such as the Designed Ankyrin Repeat Proteins (DARPins) engineered from Ankyrin repeats. As a result, DARPins are the most successful and well-studied of all engineered repeat families. Pluckthun and colleagues have proposed DARPins as alternatives to traditional antibodies (Interlandi et al., 2008, Milovnik et al., 2009). Many other reasons make engineered repeat proteins an attractive system for drug development, as DARPins have overcome the shortcomings of MAbs. DARPins have a relatively lower molecular weight (14-21 kDa) than MAbs, which can vary depending on the length and complexity of the Ankyrin domain. Owing to the considerably smaller sizes of DARPins, this, in turn, means they have a higher diffusion into peripheral tissues and organs, unlike MAbs that cannot cross the blood-brain barrier (Pardridge, 2010). Taking this further, DARPins are bound to intracellular targets, making them an extremely versatile system to engineer to bind to anything.
All repeat families and Ankyrins are widely distributed across kingdoms, meaning DARPs and other engineered repeat proteins can be expressed in any model organism or vector for mass production. This nascent ability is a direct advantage over MAbs since they depend on being expressed within eukaryotic cells to form the disulphide bonds to assemble its quaternary structure correctly and undergo post-translational glycosylation (Plückthun, 2015). Unlike naturally occurring Ankyrins, DARPs are much more stable and soluble than MAbs. These properties make them great candidates for helping to store in harsher and more remote areas that do not require expensive storing methods or facilities. Additionally, DARPs can have longer half-lives by pairing with polyethylene glycol (Stumpp et al., 2008).

With so many advantages that overcome the shortcomings of MAbs, they present an apparent system to design proteins as drugs or research tools (reviewed in Caputi and Navarra, 2020 and Stumpp et al., 2008). Many DARPs have been successfully designed as research tools (reviewed in Boersma, 2018). The knowledge of molecular cell biology has been expanded by visualising molecules in live cells to track and analyse the trafficking, localisation, functionality, and interactions of one or more biological molecules (Reviewed in Mehrotra, 2016). DARPs have been engineered to act as biosensors as well. The intracellular fluorescent sense-specific DARPin can detect either the active or inactive form of the extracellular signal-regulated kinase (ERK) (Kummer et al., 2013). This DARPin potentially serve as a tool for studying ERK function in vitro and in vivo.

The crystallisation of large macromolecular complexes, such as intrinsically disordered proteins and membrane proteins, requires a chaperone to assist the process (Lieberman et al., 2011, Bukowska and Grütter, 2013). Other DARPs have been used as crystallisation chaperones (Batyuk et al., 2016).

DARPs have been engineered to imitate various effector functions successfully that can combine multiple binding specificities and varying affinities for use in protein therapeutics.
For example, a selective tumour-killing bispecific DARPin binds to and leads to pan- Human epidermal growth factor receptor-2 (HER2) inhibitor, which is often involved with tumorigenesis (Jost et al., 2013); multivalent DARPin designed to bind to the SARS-CoV-2 spike protein, reached clinical trial phase 2 in under a year of research and development (Walser et al., 2021).

Thinking beyond DARPins, there have yet to be similar successes with other repeat families. However, taking advantage of individual repeat families that have set a foothold in specific cellular functions, mean certain repeat proteins can be quite easily designed to bind to a protein of interest within their respective interactome.

1.4.5 Repeat sequences
1.4.5.1 Long evolutionary periods mask shared ancestry

As the name suggests, a repeat in a protein sequence implies the replication of stretches of amino acids. However, similarity between evolutionarily identical repeats can fade over time, with the accumulation of point mutations and indels, rendering them eventually dissimilar (Anisimova et al., 2015). Moreover, divergent repeats remain conserved in structure, even when they are dissimilar in sequence.

The repeat MSA for the Pertactin β-solenoid protein, in Figure 1.2, illustrates how few positions (8/18) are require to fold into a conserved structure (Kajava, 2012). This protein has accumulated many point mutations, even at conserved sites. Furthermore, all repeats have indels of variable lengths between the β-strands.
1.4 Repeat families

Repeats demonstrate considerable sequence divergence among units of the same repeat family, even within a single protein. Despite the lack of conservation, building an alignment of repeats and determining the consensus sequence, as visualised in Figure 1.3: Sequence logo and conservation profile of repeat families., are often used to describe repeat families. Some repeat families, like TPR and HEAT, have the fewest conserved positions, whereas Armadillo and Ankyrin are some of the few families that have many conserved positions.
### 1.4 Repeat families

<table>
<thead>
<tr>
<th>Repeat Family</th>
<th>Sequence Logo and Conservation Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPR</td>
<td><img src="TPR.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
<tr>
<td>PPR</td>
<td><img src="PPR.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
<tr>
<td>PFTB</td>
<td><img src="PFTB.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
<tr>
<td>HEAT</td>
<td><img src="HEAT.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
<tr>
<td>Armadillo</td>
<td><img src="Armadillo.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
<tr>
<td>TAL</td>
<td><img src="TAL.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
<tr>
<td>Ankyrin</td>
<td><img src="Ankyrin.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
<tr>
<td>LRR</td>
<td><img src="LRR.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
<tr>
<td>WD40</td>
<td><img src="WD40.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
<tr>
<td>Kelch</td>
<td><img src="Kelch.png" alt="Sequence Logo and Conservation Profile" /></td>
</tr>
</tbody>
</table>

Figure 1.3: Sequence logo and conservation profile of repeat families. Belonging to Class III of repeat structure classifications (Kajava, 2012, Di Domenico et al., 2014). MSA information represent seed alignments from Pfam; Tetratricopeptide Repeat (TPR; PF13181),

---

1. Introduction 

1.4 Repeat families

27
Pentatricopeptide Repeat (PPR; PF12854), Prenyltransferase repeats (PFTB; PF00432), HEAT repeat (PF02985), Armadillo repeat (PF00514), Transcription Activator-Like Effector repeat (TAL; PF03377), Ankyrin repeat (PF00023), Leucine-Rich Repeat (LRR; PF00560); WD40 repeat (PF00400), Kelch repeat (PF01344). Each row represents annotation tracks for each repeat family visualised in Jalview. Conservation is the Zvelebil amino acid scoring scheme (Zvelebil et al., 1987) that is the default annotation track in Jalview. Consensus represents the sequence logo. Secondary structure represents the summarised SSEs in the alignment, using the DSSP definitions (Kabsch and Sander, 1983): H is an alpha-helix (Red), C is a 3/10-helix (orange), B is a beta-bridge residue (Green), E is an extended strand in a beta ladder, T is an H-bonded turn, S is a bend, and X is when both CO and NH make an H bond. Grey lines are regions without SSEs. Figure adapted from Kobe and Kajava (2000).

1.4.5.3 Multiple sequence alignments of repeats

Members of the repeats community have aligned individual repeat sequences in different ways. The most accessible and popular approach involves using existing MSAs from databases, such as Pfam; Mosavi et al. (2002) used a Pfam MSA of Ankyrin repeats, removed insertions and focused on high-occupancy columns (>50%), or canonical positions. Alternatively, Magliery and Regan (2005) filtered out non-canonical repeats from Pfam alignments of TPRs, Ankyrin repeats, C2H2-type Zinc fingers, and PDZ domains. The Intensification method (Chen et al., 2017) tackles this problem similarly by constructing a histogram of repeat lengths for a repeat family, choosing the most frequently observed length. Thus, remaining repeat annotations with identical sequence lengths are stacked to produce an MSA without using any alignment methods. Others have simply selected their choice of database(s) and aligned them with popular alignment methods covered earlier in Section 1.3.3. For example, Gul et al. (2017) used MUSCLE to align Armadillo repeats to study the Metazoan evolution of this repeat family.

There are a number of studies that build sequence-based or structure-based MSAs of all repeats present in a single protein. Choi and Weis (2005) built a structure based MSA of nine Armadillo repeats in the Plakophilin 1 protein. This was also carried out with the Ankyrin repeats of the Transient receptor potential Vanilloid (TRPV) 2 protein (McCleverty et al.,
Apart from the approaches discussed, there are no alternative alignment methods to build MSAs of repeat sequences.

1.4.6 The methodological challenges for protein repeat detection

1.4.6.1 Detecting amino acid periodicities

MSA of a protein family is the simplest and most common approach to studying a protein domain. Studying MSAs can reveal functionally important residues shared between species or between protein homologues within a single species. The conservation profile can be used to structurally categorise existing and putative proteins (Gough, Karplus, Hughey, & Chothia, 2001). However, this approach suffers with poor sequence conservation, as unconserved regions are often overlooked as having little structural or functional importance (Lichtarge, Mihalek, & Res, 2004). This is an ongoing challenge in protein bioinformatics, as many downstream methods and analytical approaches depend on meaningful alignments (Chatzou et al., 2016; Gil & Fiser, 2018).

Repeats demonstrate considerable sequence divergence among units of the same repeat family and even within a single protein. Repeat families are characterised as having a degenerate consensus sequence, which is especially true for repeats that mediate protein–protein interactions (Domenico et al., 2014; Jorda, Xue, Uversky, & Kajava, 2010). Since repeats are not so well defined in sequence, despite being near identical in structure, they are notoriously difficult to identify and annotate correctly. They are, on average, relatively short at around 60aa, or less. Furthermore, repeats are problematic when defining the boundaries (Domenico et al., 2014). With the exception of key residues that define their fold, their more diverse regions are involved in binding to substrates (Boersma & Plu, n.d.; Magliery & Regan, 2005). Current methods, such as a general Hidden Markov Model (HMM) or profiles and patterns (PROSITE), have not been able to completely capture all repeats in the way they have been applied. They have been trained on alignments of homologous or paralogous sequences
for an entire repeat family (Letunic & Bork, 2018; Sigrist et al., 2002) and, as a result of high diversity across many alignment columns, real repeats can be missed due to low scores.

There are many different methods used to study the different classes of repeats. A popular approach is to use HMM profiles, which depend on the *a priori* knowledge from an alignment of repeats. Databases such as SMART and Prosite are specialised in detecting repeat sequences. Furthermore, the HHrepID server another approach that extends from this method by comparing HMM-HMM profiles (Biegert and Söding, 2008). Other general tools, like REP2 detect eleven common repeat families (Kamel et al., 2021). In contrast, a more family specific method is the TPRpred tool that predicts regions with TPRs and two other TPR-like repeats called PPRs and Sel-1 like Repeats (SLR) (Karpenahalli et al., 2007).

Instead of HMMs, Fourier transform analysis has been used to detect arrays of repeats with fixed amino acid periodicities that do not have indels (McLachlan and Stewart, 1976, Coward and Drabløs, 1998, Gruber et al., 2005). Algorithms contrasting them are capable of tolerating irregular repeats with indels (Newman and Cooper, 2007, Jorda and Kajava, 2009). Finally, *ab initio* determination methods that carry out sequence-sequence comparison of a protein sequence with alignment methods (Heger and Holm, 2000, Szkłarczyk and Heringa, 2004).

1.5 Genetic variation

Since the HGP (Collins et al., 2003), revolutionary innovations have tackled the major barriers associated with limited throughput and the high costs of genome sequencing (NGS systems reviewed in Liu et al., 2012), which coined the familiar designation, next-generation sequencing (NGS). Advances in genomic technologies have considerably reduced the cost of sequencing to below $1000 per genome (Figure 1.4; Wetterstrand, 2021). These developments have accelerated DNA sequencing of humans, yielding large cohorts of whole-exome sequencing (WES) and whole-genome sequencing (WGS) datasets from patients and
population studies across a variety of large-scale sequencing projects (Karczewski et al., 2020a).

Figure 1.4: Cost of sequencing human genomes. Courtesy from the National Human Genome Research Institute [https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost] (Wetterstrand (2021)).

Since the internationally-led effort to clone and sequence an entire human genome in the Human Genome Project, the discovery of human genetic variation (Tishkoff and Verrelli, 2003) has led to the sequencing of hundreds of thousands of exomes and genomes. The 1000 Genomes project (1KGP) was the first catalogue of human genome sequence variation of healthy people (Consortium, 2010). The 1KGP was surpassed by the NHLBI “Grand Opportunity” Exome Sequencing Project (GO-ESP) of 6,516 exomes of European American or African American individuals (Fu et al., 2013). The Exome Aggregation Consortium (ExAC) exceeds previously available exome-wide variant databases, which consists of high-quality exome sequencing data from 60,706 individuals across diverse ancestries (Lek et al., 2016).
1. Introduction

The successor to ExAC was the Genome Aggregation Database (gnomAD) that combines all the aforementioned datasets plus 67 other large-scale sequencing projects (Karczewski et al., 2020a). GnomAD comprises 125,748 exomes and 15,708 genomes from a cohort of 141,465 healthy individuals with no known diseases. Furthermore, genetic data from unrelated individuals not including first-degree relatives, were sequenced from various disease-specific and population genetic studies from sequencing projects, such as genome-wide association study (GWAS), were aggregated and harmonised.

The main goal of organising and harmonising reams of genomes is to help understand our evolutionary history (Tishkoff and Verrelli, 2003). Efforts to study population variation encourage the development of new approaches for predicting the effects of amino acid substitution-causing variants in proteins (Adzhubei et al., 2010, Sim et al., 2012) and interpreting the effects of variants in genic (Gussow et al., 2016, Sivley et al., 2018) and non-genic regions (Vitsios et al., 2021). Instead, other databases include the catalogue of clinically validated variants, such as ClinVar (Landrum et al., 2018) and Human Gene Mutation Database (HMGD) (Stenson et al., 2020). ClinVar archives the interpretations of clinically relevant variants of any type, size, or genomic location for conditions that have been experimentally characterised. The HMGD similarly catalogues clinical interpretations of variants, but only for mutations that have been shown to be disease-associated; termed pathogenic variants. In contrast, ClinVar includes all variants, regardless of their association with disease.

1.5.1 Types genetic variation

Genetic variation in humans comprises sequence and structural changes of varying ranges of base lengths: single-nucleotide polymorphisms (SNPs) and multi-nucleotide polymorphism (MNPs), insertions or deletions (indels), copy number variation (CNV), and inversions and translocations (Reviewed in Frazer et al., 2009, Haraksingh and Snyder, 2013).
SNPs are single base pair point mutations, whereas MNPs are more than one base pair substitutions that both occur in >1% of individuals in a sampled population. Both types of nucleotide polymorphisms have no net gain or loss of genetic material.

In contrast, structural variants range in size from hundreds to hundreds of thousands, as well as entire chromosomes or even genomes. The shorter changes include single base substitutions, such as transitions and transversions, and insertion–deletions (indels) of several base pairs (Zhang et al., 2010, Choi et al., 2012). Others include block substitutions, where an array of adjacent bases are different (Levy et al., 2007). Larger changes include CNVs, which are described as a duplication or deletion event of a region of a genome that affects a considerable number of base pairs, as well as inversions and translocations (Frazer et al., 2009, Haraksingh and Snyder, 2013). Lastly, the largest genetic variation events include changes in the number of chromosomes, for example whole-genome duplication (WGD) (McGrath and Lynch, 2012) and whole-chromosome aneuploidy (WGA) of excess or deficiency in chromosomes (Giam and Rancati, 2015).

Human genetic variation occurs both within coding and non-coding regions of the genome; the majority occurring within non-coding regions of the genome (Waterston and Pachter, 2002, Hindorff et al., 2009), where SNPs can influence non-coding regulatory RNA elements (Wells et al., 2019, Minotti et al., 2018), and can sometimes lead to human disease (Zhang and Lupski, 2015). SNPs within coding regions are under higher constraint, since they can influence structure and function of proteins (Wang and Moult, 2001), such as protein stability (Yue et al., 2005), function (David et al., 2012, Yates and Sternberg, 2013), and gene expression (Hunt et al., 2014).

1.5.2 Evolution of protein-coding regions in the genome
Evolutionary constrained sites in an alignment of protein homologues provide insight into the different structural and functional pressures, as determined by amino acid conservation
1. Introduction

1.5 Genetic variation

(Barton, 1990). Likewise, global patterns of genetic variation across the human genome can also reflect the functional and evolutionary constraints at genic and nongenic regions.

SNPs within exons can come in two forms: missense and synonymous variants. Missense variants cause amino acid substitutions, while synonymous variants are silent. Early gene sequence studies have discovered a lower proportion of missense variants than synonymous variants in protein-coding regions, highlighting the high selective pressure of mutations that translate to the protein level (Cargill et al., 1999).

The genomic distribution of variants in the exome is influenced by the intolerance to genetic variation (Petrovski et al., 2013) or by gene essentiality (MacArthur et al., 2012). Delving further into subgenic regions, other factors include varying tissue-specific exon expression because of alternative splicing of messenger RNA (Cummings et al., 2020), and genomic features such as splice donor and acceptor sites (Telenti et al., 2016). Within translated protein regions, domain architecture influences the distribution of missense variants (Gussow et al., 2016). Missense variants are further spatially clustered in three-dimensional structure (Sivley et al., 2018), constrained by domain evolution. Pathogenic variants cluster at regions of functional or structural importance, whereas missense variants from reside at regions tolerant of amino acid substitutions, in the human population (Sivley et al., 2018).

1.5.3 Interpreting variants between individual protein residues

As exciting as these advances are, they are not without limitations. With Ensembl, which catalogues from dbSNP (version 146) (Sherry et al., 2001), 1KGP, ESP, and ExAC, most human sequence residues in Pfam domains have zero variants (MacGowan et al., 2017). Consequently, variant data are still too sparse to interpret variations between individual protein residues in the proteome.

For decades, MSAs have been used to interpret sequence variation at positions in evolutionarily related sequences across species (Chatzou et al., 2016). With the same logic, MacGowan et al. (2017) incorporated human population variant data into conventional
1.5 Genetic variation

sequence analysis to study sequence variation within a single species (human). Unlike other approaches covered in Section 1.5.2, this strategy presents a higher number of variants at equivalent positions in an MSA, boosting the statistical power. This approach helped understand their relationship at consensus positions of MSA across all Pfam families. Sites under selective pressure in the human population highlighted the same positions of structural and functional importance that dictate domain evolution, as observed in protein residue conservation.

1.5.3.1 Diseases and population variation associated with repeat proteins

Since many repeat proteins are involved in a wide range of molecular processes, there are several diseases associated with them. For example, the O-linked N-acetylglucosamine transferase (OGT) (O15294) which glycosylates serine/threonine residues of cytoplasmic and nuclear proteins, has several pathogenic variants that cause X-linked intellectual disability (XLID) in males, and other neurodevelopmental disorders (Pravata et al., 2020). LRR-containing proteins are prominent in the innate immune system, where mutations in the Toll-like receptor 3 protein (TLR-3) predispose individuals with impaired interferon production and reduced immune response in the central nervous system to the herpes simplex virus 1 (Zhang et al., 2007). The Armadillo-containing Adenomatous polyposis coli protein (P25054) can harbour mutations that cause cancer predisposition syndrome and other cancers associated with the digestive system (Sieber et al., 2000).

Pathogenic variants do not behave as they would normally with globular proteins. A mutation at an evolutionarily conserved position is more likely to be pathogenic than at an unconserved site. However, a tabular alignment of individual repeats behaves differently to that of a full repeat domain (Anna et al., 2021).

While the research community have characterised the pathology of mutations in repeat proteins, there have also been some attempts to understand how they contrast with population variation and how they evolve. Sousa et al. (2010) identified common genetic variants in the
LRR-transmembrane membrane 3 and LRR-neuronal 3 proteins that may be prone to autism spectrum disorders in populations of European ancestry. Schröder and Schumann (2005) reviewed Asp299Gly and Thr399Ile SNPs in the LRR domain of the TLR-4 across multiple studies. Present in approximately 10% of individuals of white ethnicity, these SNPs positively correlated with several diseases, such as Crohn’s disease, bronchiolitis in infants, and infections with Gram-negative bacteria. In contrast, individuals are protected from atherosclerosis and other related conditions. Thus, these SNPs are likely to cause decreased recognition of its ligands, producing different phenotypes.

Laskowski et al. (2016) have adopted a structural approach and integrated it with human population variation to study the WD40 domain of the TBL1-related protein 1 (Q9BZK7). The authors studied how six de novo pathogenic mutations in the TBL1XR1 gene compared with all 39 missense variants from the human population. Pathogenic mutations were generally more buried and at conserved sites, compared to population variants. Likely pathogenic variants either affected the structural fold or resided at the binding face of the WD40 domain it uses to interact with substrates.

Acknowledging the work covered in MacGowan et al. (2017), Chen et al. (2017) also recognised the variant sparsity when studying individual protein sequences, and aligned repeat sequences to increase the number of variants at equivalent sites in an MSA. Intensification is a resource for aligning repeat sequences and aggregating human variants to identify important sites constrained in the human population. The authors demonstrate their tool with the TPR protein family, and their findings are discussed in Section 3.4.5.

1.6 Scope of the thesis

The main aim of the work presented in this thesis is to propose a new and consistent workflow for studying Class III of repeat families that are specialised in substrate binding. Due to the large availability of structures and human population variation data, this offered the opportunity to investigate the relationship using an unprecedented structural and genetic
variation coverage. The opportunity to explore genetic variation in a simple structural fold that is highly degenerate on the sequence level, posed an interesting area to understand their impact on function and evolution. Additionally, many copies of repeats meant a much larger number of variants to study at equivalent sites. To take advantage of the large collection of structural data available, the ProIntVar pipeline (Protein Interactions and Variants) (MacGowan et al., 2017) was used to retrieve and collate all available structural data for each repeat family. The motive here was additionally supplemented with the focus of learning how to align repeats and establishing a new repeat and repeat architecture nomenclature for the community.

Chapter 2 explores the development of an interactive pipeline that combines repeat architecture analysis to study repeat organisation, and integrate sequence, genetic and structural data. The purpose of this is understand the relationship between human population variant data and structure and function in repeat proteins. Repeat Analysis Method (RAM) comprises a collection of tools and methods to gather sequence, genetic, and structural data to study the function and evolution of repeat families. Section 2.3.1 also discusses suggestions for how to select optimal set of annotations for downstream analyses. RAM implements a novel repeat architecture analysis that quantitatively measures the organisation and distribution of repeats in a proteome. RAM includes different alignment methods to generate an optimal MSA for downstream analyses. Lastly, it allows for the integration of two key pipelines, VarAlign and ProIntVar (MacGowan et al., 2017), to retrieve and organise genetic and structural data, respectively, to infer and support sites of importance, respectively.

Chapters 3 to 5 focuses on the overall genetic and structural analysis of the three TPR, Ankyrin, and Armadillo repeat families. Each chapter encompasses their contribution to RAM and its development. Each repeat family raised new questions that RAM could not answer at the time. RAM inferred previously unobserved positions of importance, that conventional sequence analysis did not answer. All the while, the depth of structural data to study the
structural and functional roles were dissected in each repeat family. The variant analysis was further extended by understanding the physicochemical properties of missense mutations. The structural data provided the opportunity to identify the structural roles of each position, by identifying positions with the number of contacts.
Chapter 2
Development of RAM

2.1 Preface

This chapter overviews the general methods and details their implementation in studying a protein repeat family. The main computational pipeline, the Repeat Analysis Method (RAM), integrates the approaches employed here to learn about repeat families specialised in protein-protein interactions. Because the development of RAM underpins all the studies on the individual protein repeat families discussed in the other chapters, this chapter describes the general repeat family analysis. The various methods and datasets introduced throughout this chapter will be revisited in more detail in the relevant chapters describing the individual repeat families.

Some components in RAM have been co-developed with Javier S. Utgés, whom I co-supervised when he was a Master’s student. His contributions were integrated into RAM and are discussed here. Javier first developed an algorithm that resembles conventional realignment that automates over groups of repeat sequences with of a certain length. Lastly, he introduced
2. Development of RAM

2.2 Introduction

the approach to study the residue surface accessibility and SSE assignment across all structures for sequences in the MSA.

Other resources were developed by other members of the Barton group: the VarAlign pipeline was established by Stuart S. MacGowan, which encapsulates ProIntVar written by Fábio Madeira (MacGowan et al., 2017).

2.2 Introduction

The motive for developing a universal method for studying repeats in proteins was the lack of consistency in the research community studying them. Initially, Kobe and Kajava (2000) defined solenoid domains in terms of their geometric properties, the various structural units observed, and the consensus sequence defining the various families that meet the criteria of solenoid proteins. Furthermore, Kajava (2012) followed up with a rigorous classification of repeats into different groups. However, beyond these classifications, there are no clear guidelines or a consensus on how the research community should study repeats. As a result, researchers have applied different approaches that best suited individual families.

It is challenging to select the most appropriate set of annotations when databases have different annotation rulings. Identifying repeats remains a challenge, particularly when detecting repeats with indels or many point mutations that have gradually accumulated over long evolutionary times (Kajava, 2012, Anisimova et al., 2015). However, some databases are better at annotating repeats than others, and individual studies propose new methods to overcome the fragmented datasets of repeat annotations. In addition, there is inconsistent use of terminology and language when referring to, or describing, repeats.

As introduced in Section 1.4.5.1, repeats are more conserved in structure than in sequence and so, as expected, repeat sequences are challenging to align correctly. Thus, many studies that conduct repeat family analysis avoid the alignment stage and only select repeats representing the most common length for sequence analysis.
Despite a growing number of structures being solved, the community is doing very little to incorporate the extensive collection of structural data in the PDBe. As a result, studies often represent their findings using selected structures rather than conducting a comprehensive structural analysis.

These are three critical areas lacking clarity and requiring solutions to overcome the challenges and address inconsistencies in studying repeat families. The following section discusses the hierarchy of this new workflow using RAM and explains how it solves these problems.

2.3 Methods and Contents

As discussed in Section 1.4.3, different repeat families share core similarities. In reviewing many studies, they share some common features in the strategies employed to address the gaps in knowledge of each family. The hierarchy can be broken down into five steps as presented in Figure 2.1:

1. Gathering repeat annotations
2. Repeat architecture
3. Multiple sequence alignment
4. Human population variation
5. Structural analysis
2. Development of RAM

2.3 Methods and Contents

Figure 2.1 Schematic representation of the RAM protocol. RAM works in five main stages: 1) Gathering repeat annotations from InterPro and carrying out a non-redundant merge of all repeat annotations from different contributing member databases; 2) Protein repeat architecture summarises the sequential organisation of the repeats; 3) build a multiple sequence alignment, calculate a tree with appropriate clustering method, and carry out sequence analysis of the whole alignment and the subfamilies; 4) integrate human population variant data to the MSA, classifying alignment columns that are relatively enriched or depleted for missense variants, and infer positions of interest; 5) conduct a quantitative structural analysis with contact maps across all available three-dimensional structures from the PDBe and identify key positions that are constrained. The example provided is for the Armadillo repeat family (IPR000225).

In step 1, gathering repeat annotations entails researching which databases annotate the repeat family of interest and understanding what is known from the literature. Once the databases of interest have been identified, RAM carries out a non-redundant merge of the relevant contributing member databases that best represent the repeats.

Step 2 in the workflow involves studying how the repeats arrange themselves in the primary sequence across all proteins. Step 3 involves building a high-quality MSA using an appropriate method and conducting an in-depth sequence analysis to identify evolutionarily conserved positions in the MSA. Sequence analysis is further expanded by identifying subfamilies based on different properties of the repeat family. Step 4 integrates human population variant data into the MSA and identifies the selective pressures of each position in the population. The last step, 5, retrieves all available structures from the PDBe for a
comprehensive structural analysis. Steps 4 and 5 loop together to understand how structure influences the conservation and variation features in step 5, and how those features are constrained by structure in step 4, so they can be used to predict or contextualise structural features.

The common strategies applied here will be combined into a workflow that proposes a new approach to conducting a proteome-wide repeat family analysis. In addition, it is designed with a modular architecture to incorporate improved methods and new data in the future.

2.3.1 Gathering repeat annotations

2.3.1.1 Reviewing the literature

The literature is a starting point to look for repeat annotations and identify how the community defined the repeat. First, explore the structural motif for the repeat of interest by sampling different structures from the PDBe for reviewed proteins in UniProtKB/Swiss-Prot. Then decide if it is a suitable repeating monomer when viewing it structurally with adjacent repeats in a solenoid domain. Figure 2.2A shows how different databases have different structural motif definitions for a repeat, including clashing annotations from other repeat families. For example, the three underlying databases annotating Armadillo repeats in the Importin subunit alpha-8 (C1JZ66) protein show that PROSITE (PS50176) differs from Pfam (PF00514) and SMART (SM00185) in the annotation ruling by labelling the three helices as H3, H1, and H2. The different structural motif definition for PROSITE is likely due to the detection of the repeating sequence when the structure did not exist at its generation. Thus, the structural motif is offset by one helix.

Furthermore, another repeat family can annotate the same region. The Armadillo repeat family is evolutionarily related to HEAT repeats (Andrade et al., 2001). HEAT (IPR004155) annotations in the same regions as Armadillo repeats add another layer of complexity when deciding what repeat family this protein belongs to, especially if no structures are available.
2. Development of RAM

2.3 Methods and Contents
Figure 2.2 Entry matches for three repeat-containing proteins in InterPro. (A) Importin subunit alpha-8 (C1JZ66, IMA8_BOVIN) protein from *Bos taurus* (Bovine) and all the annotations. (B) Intraflagellar transport protein 56 (A0AVF1, IFT56_HUMAN) protein from *Homo sapiens*. (C) 26S proteasome non-ATPase regulatory subunit 10 (O75832, PSD10_HUMAN) protein from *Homo sapiens*. The proteins have four groups of annotation tracks for each biological entity; Family, Domain, Homologous Superfamilies, Repeats. Other annotation tracks includes: Unintegrated, other annotations that were not included; Other Features, are MobiDB Lite that provides consensus predictions of intrinsic disorder; Predicted 3D structures and Domains, are annotations of predicted three-dimensional structures and domains; Match Conservation, is the Pfam conservation score. The grey bar top of the figure represents the length of the protein with a ladder of amino acid intervals. Each group of annotation tracks represent InterPro Entries and its member databases. Screenshots from InterPro webpage (Finn et al., 2017).

2.3.1.2 Assess dataset quality

The next stage in the workflow determines which databases have annotated the repeat family of interest and analyses how well they have achieved this. This section will discuss the important databases from which to source repeat annotations and how best to select them with the tools in the workflow. Additionally, the challenges and difficulties involved with selecting them will be covered here.

2.3.1.2.1 Collecting and organising annotations

RAM uses InterPro and UniProtKB/Swiss-Prot databases as the primary and default sources of annotations but can include datasets from other databases and methods. All UniProtKB proteins, InterPro entries, and individual signatures they match for the InterPro release version 75.0 (Table 2.1), in XML format, were downloaded from the FTP server with GNU Wget software. This dataset was parsed and converted into a CSV file format to work with the Python Pandas library. A TSV format of UniProtKB/Swiss-Prot (release version 2020_06) was downloaded via the API. This dataset comprises 136212 reviewed proteins across 13984 species.
2. Development of RAM

2.3 Methods and Contents

Table 2.1: Contributing member database information for InterPro release version 75.0. There are 13 signature databases integrated into InterPro.

<table>
<thead>
<tr>
<th>Signature Database</th>
<th>Version</th>
<th>Signatures</th>
<th>Integrated Signatures</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATH-Gene3D</td>
<td>4.2.0</td>
<td>6119</td>
<td>2446 (40%)</td>
</tr>
<tr>
<td>HAMAP</td>
<td>2019_01</td>
<td>2274</td>
<td>2272 (99.9%)</td>
</tr>
<tr>
<td>Pfam</td>
<td>32.0</td>
<td>17929</td>
<td>17420 (97.2%)</td>
</tr>
<tr>
<td>PIRSF</td>
<td>3.02</td>
<td>3285</td>
<td>3217 (97.9%)</td>
</tr>
<tr>
<td>PRINTS</td>
<td>42.0</td>
<td>2106</td>
<td>1953 (92.7%)</td>
</tr>
<tr>
<td>PROSITE</td>
<td>2019_01</td>
<td>1310</td>
<td>1287 (98.2%)</td>
</tr>
<tr>
<td>PROSITE</td>
<td>2019_01</td>
<td>1232</td>
<td>1173 (95.2%)</td>
</tr>
<tr>
<td>PANTHER</td>
<td>14.1</td>
<td>123151</td>
<td>9234 (7.5%)</td>
</tr>
<tr>
<td>SFLD</td>
<td>4</td>
<td>303</td>
<td>147 (48.5%)</td>
</tr>
<tr>
<td>SMART</td>
<td>7.1</td>
<td>1312</td>
<td>1264 (96.3%)</td>
</tr>
<tr>
<td>SUPERFAMILY</td>
<td>1.75</td>
<td>2019</td>
<td>1601 (79.3%)</td>
</tr>
<tr>
<td>TIGRFAMs</td>
<td>15.0</td>
<td>4488</td>
<td>4435 (98.8%)</td>
</tr>
<tr>
<td>CDD</td>
<td>3.17</td>
<td>14908</td>
<td>3084 (20.7%)</td>
</tr>
</tbody>
</table>

Signature Database: Name of database. Version: Release version of the signature database. Signatures: Total number of protein signatures in the signature database for that release version. The absolute number and proportion (in percentage) of protein signatures incorporated into InterPro.

2.3.1.2.2 Identifying appropriate InterPro Repeat Entry

Figure 2.2C demonstrates how InterPro functionally annotates this Ankyrin-containing protein for different Biological Entities. This protein has a well-annotated set of contiguous Ankyrin Repeat annotations. The Family Entities are not helpful when studying repeats; they describe groups of proteins with a common evolutionary origin that often refer to complete proteins rather than repeats or domains. Annotations on the Domain level can span multiple repeats, for example InterPro annotates two Ankyrin repeats (IPR002110) in Figure 2.2 with InterPro Entry IPR020683. Lastly, Homologous superfamily Entities label regions very likely to have repeats, which will be discussed later in Section 2.3.6. Therefore, the first, and manual, step of RAM involves understanding what Biological Entities best represent repeats and how to select the optimal dataset for downstream analysis.
When describing repeat annotations for Repeat Entities, several terms are established here: *coverage* implies the number of proteins annotated for a given InterPro Entry; *canonical* repeats have sequence homology with most other repeats in the family and represent the most common length, and therefore a repeat with significantly large indels can be marked as a *non-canonical* repeat; *comprehensive* refers to the prevalence of gaps between repeats comparable to a *canonical* repeat.

All annotations in Repeat Entities are assessed first. The protein signatures best suited for a protein repeat family analysis should have high coverage on the Repeat Entity level since this analysis requires canonical repeats.

### 2.3.1.2.3 Study discrepancy between databases

RAM can study the relationship between the selected InterPro Entries and captures which regions are annotated by multiple databases and which are unique. This algorithm organises repeat annotations from different protein signatures into groups that label similar regions. Inspired by InterPro using the Jaccard similarity index and containment index to evaluate if a pair of annotations belonging to the homologous superfamily and InterPro entry are related to each other (Mitchell et al., 2019), the same method was employed here. The algorithm in RAM obtains the start and end positions for both annotations A and B and measures their overlap with a Jaccard Index (JI) (Jaccard, 1912) defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

The formula for JI (Jaccard, 1912) entails the intersection (coloured red in Figure 2.3A) between annotation A and B (|A\cap B|). The intersection is divided by the union (coloured teal in Figure 2.3A) of annotation A and B (|A\cup B|). The algorithm calculates all unique pair combinations within each protein. Two annotations are defined as similar if they meet the threshold $jT$. The paired annotation data was used as input to build a disjoint-set data structure
(Galler and Fisher, 1964) and categorise all annotations. The process iteratively adds new sets, merges existing sets, and determines whether annotations are in the same set. The algorithm identifies three sets of annotations in the example of Figure 2.3A, each labelling a specific region of this hypothetical protein.

Demonstrating this algorithm with a real dataset, the UpSet plot (Lex et al., 2014) in Figure 2.3B visualises the relationship between all the databases and its annotations for the TPR family. Unlike Venn diagrams (Venn, 1880), UpSet plots provide an alternative way visualise set overlaps, which shows the number of repeats each database annotates here. For example, databases PROSITE (PS50005), SMART (SM00028), and UniProtKB/Swiss-Prot (TPR) share the intersection size of 388 TPR annotations in the vertical bar plot. In contrast, the SMART database annotates 28 unique TPR annotations not recognised by other databases. Additionally, Figure 2.3B demonstrates the discrepancy between the databases annotating a single repeat family. Combining all relevant annotations is recommended since one protein signature cannot capture all repeats within a family.
2. Development of RAM

2.3 Methods and Contents

Figure 2.3: Grouping similar repeat annotations with TPRs. (A) A hypothetical protein with six annotations labelled A to F. Each row represents annotations labelled by a member database numbered 1 to 3. Annotations are coloured based on what set they belong to after pairwise comparisons, where each pair of annotations within each set meet the jT threshold. The N-to-C black line indicates the primary sequence of the protein. Calculating JI involves the intersection (red) with annotations A (Database 1) and B (Database 2) to be divided by the union (teal) of them. Sets 1 to 3 indicate the collection of annotations that annotate the same regions. (B) The Upset plot of TPR annotations from member database signatures that contribute to the InterPro Entry (IPR019734) and from Swiss-Prot (‘TPR’) for human proteins. Set size refers to the total number of TPR annotations detected by each member database visualised in the horizontal bar plot; PROSITE (PS50005) annotates 519 TPRs, and SMART (SM00028) annotated 649 TPRs. The vertical bar plot shows the number of TPR annotations within each set overlap, where the contributing sets are indicated with a black dot and vice versa in grey. For example, 128 TPRs are annotated by both SMART and Swiss-Prot, whereas PROSITE and SMART annotate 28 TPRs.

2.3.1.2.4 Selecting appropriate databases

The workflow permits a non-redundant merge of all the annotations to create a comprehensive dataset. This step initially involves selecting the order of databases to merge,
based on analysing the literature and annotation mapping. Using the example in Figure 2.3B, an order of SMART, PROSITE, Pfam, and Swiss-Prot produces a non-redundant dataset of 649 SMART, 346 Swiss-Prot, and 33 PROSITE annotations. This step generates 1028 annotations across 166 TPR-containing proteins to build an MSA, as described in Section 2.3.3.

The last step can be deemed unimportant since InterPro already provides non-redundant annotations for all Entries. However, InterPro often incorrectly merges Repeat Entity annotations from contributing member databases with different repeat annotation rulings. As a result, specific repeats are artificially long and span two repeats. For example, two InterPro Armadillo repeats are unusually long in the Armadillo-containing protein of Figure 2.2. These long Armadillo repeats result from InterPro merging annotations across databases with different annotation rulings into a single repeat. Therefore, RAM rebuilds annotations from member databases of choice.

Selecting the order of merging of databases is vital. First, the profile-based databases integrated into InterPro have a consistent annotation ruling. These would be favoured first over Swiss-Prot database. Swiss-Prot can incorrectly annotate repeats in reviewed protein entries for multiple reasons. The curator(s) for an entry may have accepted a set of annotations incorrectly referenced from a study, or defined differently within the study itself. Often no evidence explains the set of annotations in the protein entry. In some cases, UniProt annotates an evolutionarily related repeat family with sequence and structural homology to the correct repeat family. For example, Zinc finger C3H1 domain-containing protein (O60293) and DAP3-binding cell death enhancer 1 (Q14154) protein belong to a small set of proteins having TPR-like (IPR003107) annotations in InterPro and TPR annotations in Swiss-Prot.

Another reason to favour selected profile-based databases is that the Swiss-Prot curator can incorrectly search a protein with a sequence profile from a database that defines the wrong structural motif. For example, the General vesicular transport factor p115 protein (P41541) has
its Armadillo repeats annotated with the PROSITE ruling H3-H1-H2, whereas SMART and Pfam annotate correctly with the order H1-H2-H3. These clashing rulings do not come as a surprise since PROSITE may have originally built profiles from sequence alone before structures were available for certain protein families.

A further reason is the advantage of selecting the order of databases that favour specific repeat annotations labelling the whole structural motif. For example, Pfam annotates an incomplete length of an Armadillo repeat, unlike SMART. This example may be because Armadillo repeats are more conserved in their helices H2 and H3, as explained in the Armadillo repeat family analysis of Chapter 5.

Despite the issues mentioned above, the advantage of Swiss-Prot is the manual curation of repeats that are unusually long or short because of indels. In addition, structures remain the gold standard for annotating functional regions of proteins and thus, help determine the non-canonical repeats that remain undetected by profile-based methods.

2.3.2 Repeat architecture

PDA describes the linear organisation of functional domains in proteins. Characterising how repeats arrange themselves into solenoid domains is no different from PDA. Protein repeat architecture (PRA) is proposed here, which denotes the organisation of repeats within domains and the repeat domains themselves in primary sequence.

This section will discuss the new terminologies to characterise PRA and refer to repeats. In addition, the new language will help formulate new questions that this method can address. Finally, this method will demonstrate how dissecting the organisation of repeats with this new and sophisticated approach can bring new insights into the function and evolution of repeat families.
2.3.2.1 Redefining repeat nomenclature

There is inconsistent use of language to describe repeats in the literature. Motif and Repeat are often used interchangeably or in conjunction as Repeat motif (D’Andrea and Regan, 2003, Main et al., 2005, Zeytuni and Zarivach, 2012, Andrade et al., 2001). Lamb et al. (1995) termed the individual helices of a TPR as domains. Such descriptions are observed in earlier literature, before different databases had defined repeats as a distinct class that differ from domains and motifs. Although clarified clearly as a term, Bjorklund et al. (2006) conceived new terminologies to refer to repeats as short repeating domains or Domain Repeats. Perez-Riba and Itzhaki (2019) characterised the TPR domain of Peroxisomal targeting signal 1 receptor protein, with seven adjacent TPRs, as two TPR subdomains separated by the fourth TPR. This TPR acts as a hinge in the middle of the two TPR subdomains to help fold the TPR domain into a clamp (Stanley et al., 2007, Stanley et al., 2006). The repeat research community frequently use the word repeat in conjunction with the abbreviated form where it already is an acronym, coined as Redundant Acronym Syndrome (RAS) (RAS reference). For example, Leucine-Rich Repeat (LRR) and Tetratricopeptide Repeat (TPR) are written as an LRR repeat and TPR repeat, respectively (Enkhbayar et al., 2004, Cortajarena et al., 2004).

A new nomenclature is proposed here to tackle the discrepancy in the literature and help refer to, and describe, repeat families for the novel analyses discussed later in this chapter. There are two sets of terminologies that characterise repeats: internally within a family, and globally when referring to similar repeat architecture features across multiple families in general. To demonstrate this new language, Figure 2.4 visualises the terminologies in the TPR family. A TPR is a single helix-loop-helix-loop structural motif (Section 3.2). The spacing $S$ between two TPRs and threshold $T$ is measured in the number of amino acids. A TPR-Domain is defined as two or more adjacent TPRs that meet the $S \leq T$ condition. Therefore, a TPR-Domain with two or more TPRs is called 2-TPR-Domain or $N$-TPR-Domain, where $N$ is any
number larger than 2. A TPR in a TPR-Domain is specified by including its numerical position within the TPR-Domain term. For example, in Figure 2.4, the second TPR in a 3-TPR-Domain is denoted as a 3-TPR-2-Domain or 3-TPR-2 for conciseness. These new terminologies were designed to extend to other repeat families, where TPR in this example could be an Armadillo, Ankyrin, and any other repeat family.
Figure 2.4: Schematic representation of repeat architecture demonstrated with the TPR family. Individual orange stadium-shaped objects represent a TPR, and the black lines denote the primary sequence. (A) Spacing $S$ denotes the gap, in amino acids, between any two TPRs. Threshold $T$ applies to the spacing, where TPRs are classified as adjacent when they meet the...
2. Development of RAM

S ≤ T condition. Repeat architecture labels all TPRs relative to other repeats on different levels; the TPR position in the protein, the TPR position in the TPR-Domain, and the TPR-Domain position. Referring to individual TPRs are demonstrated with 2- and 3-TPR-Domains. (B) The PRA for three examples of TPR-containing proteins with annotations from InterPro (IPR019734).

The global terms are similar in form to the internal ones. Like a TPR-Domain, a Repeat-Domain is a general term referring to adjacent repeats that fold into a solenoid domain. A 3-Repeat-Domain denotes any Repeat-Domain comprising three adjacent repeats. All Repeat-Domains can be further broken down to describe capping repeats on the N- and C- terminal ends, and all the other repeats in between. The N-terminus end is a start repeat, and the C-terminus end is an end repeat. 3-Repeat-Domains or longer also have internal repeats between start and end repeats, termed middle repeats.

2.3.2.2 Repeat architecture analysis

Repeat architecture analysis entails visualising the high-level and proteome-wide organisation of a repeat family, shown using TPRs as an example in Figure 2.5. The first part of the analysis visualises the spacing frequency between repeat annotations. The purpose of this is twofold: to identify gaps that are multiples of N, where N refers to the length of a canonical repeat; and to determine the spacing threshold when grouping adjacent repeats into domains. To illustrate this decision-making process, Figure 2.5A shows the spacing frequency for a non-redundant set of 1028 TPR annotations produced using methods described in Section 3.3.3.2. Canonical TPRs are 34 amino acids long. Studying the spacing frequency for multiples of 34aa with ±5aa range due to indels, there are very few gaps that resemble a TPR. For a dataset with many missing repeats, a different observation with peaks at those regions would indicate the presence of potentially missing TPR annotations. In addition, the spacing frequency plot will highlight if the dataset is fragmented, as discussed earlier in Section 2.3.1.2, thus requiring a better set of annotations, or reannotation of repeats using the maxHMM algorithm, as summarised in Section 2.3.6.
2. Development of RAM

2.3 Methods and Contents

Figure 2.5: Repeat architecture analysis for the TPR family. (A) Spacing frequency shows the number of spacings between any two TPRs. (B) The total number of TPRs. (C) The total number of TPR-N-Domains, where N indicates the number of TPRs in a TPR-Domain. For example, there are eight 7-TPR-Domains in this dataset. (D) The total number of TPR-Domains in each protein. For example, there are 32 proteins that have two TPR-Domains, regardless of
how many TPRs they have. (E) Heatmap showing the combinations of different Armadillo-Domains and how many proteins display these patterns. For example, in row F, six proteins have one 3-TPR-Domain.

![Heatmap showing combinations of different Armadillo-Domains and the number of proteins displaying these patterns.](image)
The second purpose of the spacing frequency plot dictates the output of the rest of the repeat architecture analysis. Applying a threshold of half the length of a canonical repeat is an arbitrary decision that is a default parameter in RAM. Nevertheless, the shortest repeat that folds into the conserved structural motif representative of the repeat family would be recommended. For example, short Armadillo repeats tend to be the capping repeats with two helices instead of three, whereas middle Armadillo repeats tend to have insertions rather than deletions. For example, Figure 2.6 shows how these two Armadillo repeats preserve the three-helical fold of the structural motif.

Figure 2.6: Structural alignment of Armadillo repeats and MSA. The three Armadillo repeat structures were aligned with STAMP (Russell and Barton, 1992) and its resulting MSA. Ribbon representation of structures was visualised with UCSF Chimera (Pettersen et al., 2004). Jalview benchwork (Waterhouse et al., 2009) visualised the MSA with the ClustalX colour scheme and the annotation track. Conservation scores the sequence conservation (Zvelebil et al., 1987). (A) Structural alignment of three Armadillo repeats: canonical Armadillo repeat from Vacuolar protein 8 (VAC8_YEAST, P39968) in orange (PDB code: 6KBM), non-canonical Armadillo repeat from the General vesicular transport factor p115 protein (USO1_BOVIN, P41541) in blue (PDB code: 3GRL) that has an insertion between H1 and H2,
non-canonical Armadillo repeat from the Junction plakoglobin protein (PLAK_HUMAN, P14923) in green (PDB code: 3IFQ) that has an insertion between H2 and H3.

The following plot in Figure 2.5B shows the total number of TPRs, regardless of their position in the primary sequence level. Figure 2.5B is reminiscent of the PRA analysis of TPRs by D'Andrea and Regan (2003). The PRA analysis next groups adjacent TPR annotations, with a spacing threshold of 20aa, into TPR-Domains in Figure 2.5C. This plot is informative as it visualises all TPR-Domains used by proteins and shows the evolutionarily functional Repeat-Domains in this family. Repeat-containing proteins can have more than one Repeat-Domain, permitting such proteins to bind to multiple substrates. For example, Stress-induced-phosphoprotein 1 is capable of binding both Heat Shock Protein 70 (HSP70) and 90 (HSP90) with the N-terminus and C-terminus TPR-Domains, respectively (Chen et al., 1996, Lässle et al., 1997). To capture this information, Figure 2.5D visualises the number of TPR-Domains, regardless of the number of TPR comprising them.

The final element of this analysis combines all the high-level PRA data to visualise how proteins contribute to each element of PRA. Figure 2.5D visualises this information to understand how different TPR-Domains organise themselves in proteins. However, one element of the repeat architecture analysis not captured in Figure 2.5D is that it disregards the sequential order of Repeat-Domains in a protein. Instead, this analysis captures the different Repeat-Domains as independently functioning Repeat-Domains. Therefore, the five TPR-containing proteins in Row M of Figure 2.5D may be organised as 3-TPR-Domain and 5-TPR-Domain in the N- to C-terminus direction or vice versa.

The main caveat of this method is the presence of large insertions between structurally adjacent repeats that preserve a contiguous solenoid structure. Figure 2.7 shows how a large insertion, 77aa (615–692), does not interfere with the TPR-Domain fold. Instead, the PRA analysis would determine them as two separate TPR-Domains without a structure. Despite this
example, the analysis is generally valid since these situations are rare and it is impossible to quantify this at the time of writing, as there are far fewer structures than sequences available.
Figure 2.7: AlphaFold model of Tetratricopeptide repeat protein 7B. Ribbon representation of the region 431-843aa of TTC7B_HUMAN (Q86TV6) was visualised with UCSF Chimera. Coloured according to the model confidence from AlphaFold (Jumper et al., 2021): very high (pLDDT > 90), in dark blue; confident (90 > pLDDT > 70), in light blue; low (70 > pLDDT > 50), in yellow; very low (pLDDT < 50), in red. Structure visualised in UCSF Chimera.
2.3.3 Multiple sequence alignment

Introduced in Section 1.4.5, repeat sequences have a few positions that must be conserved across the family to preserve the structural fold. With this distinct feature, repeats remain a challenge to align correctly. While, a variety of studies have circumvented this problem using different approaches, two alternative strategies have been implemented into RAM to address the difficulty of aligning repeats. These improve existing methods and approaches by successfully aligning repeats with indels and, thus, include non-canonical repeats.

2.3.3.1 Conventional alignment with realignment

RAM can employ most popular and standard global alignment methods: Clustal W (Larkin et al., 2007); MAFFT (Rozewicki et al., 2019); T-Coffee (Notredame et al., 2000); MSAProbs (Liu et al., 2010); ProbCons (Do et al., 2005); Clustal Omega (Sievers et al.); GLProbs (Ye et al., 2014); MUSCLE (Edgar, 2004).

The conventional approach to aligning sequence entails adjusting and optimising different gap penalty parameters. When building MSAs, there tends to be a subset of sequences that make up many insertions introduced by the alignment method. Correcting for gap-occupying sequences and refining the final MSA involves realigning those sequences to the MSA using the same alignment method previously used. Repeat families have more copies than most protein domains, requiring many realignment stages. Instead of manually realigning, Javier S. Utgés devised an algorithm to automate this for repeats, and applied the conventional alignment approach to specific groups of repeats.
2. Development of RAM

2.3 Methods and Contents

Figure 2.8: Recursive alignment algorithm. (1) Align first sequence group with ClustalΩ. (2) Dealign sequences that occupy insertions in the MSA. (3) Realign those sequences up to three times to the main MSA. (4) Sequences that failed to realign after three attempts are set aside. (5) Select next sequence group and restart loop at step 1 by aligning the sequences from the group to the growing alignment with a sequence-to-profile alignment approach.

With a priori knowledge of the repeat length and the databases, the algorithm categorises repeats into sequence groups (SGs). For example, the first SG for the Ankyrin repeat family was the canonical repeat length of 33aa from PROSITE. SGs are organised in
descending order according to sequence length across the different databases for the same reasons discussed in Section 2.3.1.2.4. The algorithm aligns SGs with shorter sequence lengths before addressing repeats longer than the canonical length. The purpose of the sequence order is to avoid the alignment of shorter sequences to columns introduced in the MSA by ClustalΩ because of longer sequences.

Step 1 in Figure 2.8 is aligning the SG with ClustalΩ. If the resulting MSA had gaps at high-occupancy columns, with more than 50% residue occupancy, the alignment algorithm realigned gap-occupying sequences in step 2 (Figure 2.8). In step 3, the algorithm realigns those sequences to the MSA with a sequences-to-profile approach. Sequences that still introduce gaps after three realignments to the MSA, they were temporarily set aside in step 4. The continued introduction of gaps in the MSA by the sequences throughout the realignment stages can be interpreted as genuine indels of those sequences. In step 1’, the loop restarts and the algorithm selects the next SG and aligns to the growing MSA with a sequence-to-profile alignment.

After all SGs were processed, only 2% of the 7,407 were set aside as ClustalΩ kept introducing insertions in the main MSA. These sequences were divided into SGs again. Each SG was aligned to a reference MSA of \( \sim 3,200 \) canonical length sequences with no gaps. The resulting alignments were sequentially aligned to the main alignment with a profile-to-profile alignment. After aligning each SG, the reference MSA was removed.

2.3.3.2 Secondary-structure-based gap penalties

The second approach aligns sequences to a template MSA with secondary-structure-dependent gap penalties using the AMPS suite of programs (Barton, 1994). Since repeats are more conserved in structure, there should be much greater confidence in aligning sequences to a structural alignment or high-quality MSA. Therefore this strategy avoids the introduction of
insertions within regions that have SSEs in the MSA due to diverse repeats that may have indels.

A structure-based alignment of repeats can act as a template MSA. For example, a structural alignment of three Armadillo repeats with the program STAMP (Russell and Barton, 1992) is demonstrated in Figure 2.6. More accessible alternatives are high-quality MSAs, such as the seed or full alignments from Pfam. Since SSEs do not generally tolerate indels, the MULTALIGN program in AMPS incorporates this feature by implementing variable gap penalties. The default parameters in the RAM method includes a gap penalty of 10 for regions with no secondary structure elements. The gap factor of 100 multiplies the standard gap penalty within the column ranges defining SSEs. The resulting MSA in Figure 2.9B shows how all the insertions reside between the SSEs.
Development of RAM

2.3 Methods and Contents

Figure 2.9: Template MSA and output MSA from AMPS for Armadillo repeats. (A) Template MSA from Pfam (PF00514). An average distance tree was calculated for both MSAs in Jalview with BLOSUM62 substitution matrix, sorted by the resulting dendrogram. The sequences
shown for both MSAs represent equidistant members of the dendrogram. Alignments were viewed in Jalview.

2.3.4 Evaluating the quality of a multiple sequence alignment

When building an MSA for any repeat family, it is still essential to assess the performance of multiple methods. Sections 2.3.3.1 and 2.3.3.2 performed better than most standard and popular methods when aligning TPRs, Ankyrins, and Armadillos. A consistent framework to quantitatively assess the quality of an MSA was implemented into RAM.

Three parameters can be captured from all MSAs under assessment: the number of sequences that contribute to alignment columns with 50% or more residue occupancy in the MSA; the number of sequences that have at least one residue contributing to alignment columns with less than 50% residue occupancy in the MSA, and this subset of sequences can be split further by their lengths; the number of alignment columns that have less than 50% residue occupancy in the alignment.

Table 2.2 shows how different alignment methods performed when aligning sequences of 2270 Armadillo repeat sequences from Chapter 5. The global alignment of Armadillo repeats with methods listed in Table 2.2, with default parameters, gave mixed results. The two values focused on here are gap-free and gap-occupying sequences. The ascending order of improving performance using the two metrics was: T-Coffee (Notredame et al., 2000); ClustalΩ (Sievers et al.); Clustal W (Larkin et al., 2007); MUSCLE (Edgar, 2004); MAFFT (Rozewicki et al., 2019); ProbCons (Do et al., 2005); MSAProbs (Liu et al., 2010); GLProbs (Ye et al., 2014); and AMPS (building on a template MSA). T-Coffee and ClustalΩ, the worst performers, both introduced a gap in all the sequences. In comparison, AMPS, the best alignment method, had the highest possible number of gap-free sequences and the fewest gap-occupying sequences. Furthermore, AMPS also had the fewest low occupancy columns. Of the gap-occupying sequences, RAM further dissected them by the lengths of Armadillo repeats making up the
low-occupancy columns (residue occupancy < 50%). All Armadillo repeats longer than canonical Armadillo repeats contribute to this number (not shown in Table 2.2), which stands true for all alignment methods.

Table 2.2: The performance of nine different global alignment methods assessed on 2270 Armadillo repeat sequences. Each alignment program used here had default parameters. Alignment programs have been sorted in descending order of performance by the number of gap-free sequences generated.

<table>
<thead>
<tr>
<th>Sequence Alignment Program</th>
<th>Gap-Free Sequences</th>
<th>Gap-Occupying Sequences</th>
<th>Low Occupancy Columns</th>
<th>Short Length Sequences</th>
<th>Canonical Length Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMPS</td>
<td>1140</td>
<td>1130</td>
<td>55</td>
<td>86</td>
<td>106</td>
</tr>
<tr>
<td>GLProbs</td>
<td>1085</td>
<td>1185</td>
<td>463</td>
<td>124</td>
<td>123</td>
</tr>
<tr>
<td>MSAProbs</td>
<td>1084</td>
<td>1186</td>
<td>482</td>
<td>116</td>
<td>132</td>
</tr>
<tr>
<td>ProbCons</td>
<td>1078</td>
<td>1192</td>
<td>465</td>
<td>136</td>
<td>118</td>
</tr>
<tr>
<td>MAFFT</td>
<td>209</td>
<td>2061</td>
<td>378</td>
<td>433</td>
<td>690</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>23</td>
<td>2247</td>
<td>397</td>
<td>456</td>
<td>853</td>
</tr>
<tr>
<td>ClustalW</td>
<td>18</td>
<td>2252</td>
<td>110</td>
<td>467</td>
<td>847</td>
</tr>
<tr>
<td>ClustalΩ</td>
<td>0</td>
<td>2270</td>
<td>243</td>
<td>479</td>
<td>853</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>0</td>
<td>2270</td>
<td>596</td>
<td>479</td>
<td>853</td>
</tr>
</tbody>
</table>

Sequence Alignment Program: The name of the alignment program. Gap-Free Sequence(s): The number of sequences that solely contribute to alignment columns with 50% or more residue occupancy in the alignment. Gap-Occupying Sequence(s): The number of sequences with at least one residue contributing to alignment columns with less than 50% residue occupancy in the alignment. Low-Occupancy Columns: The number of alignment columns with less than 50% residue occupancy in the alignment. Short Length Sequences: The number of gap-occupying sequences shorter than the canonical length of Armadillo repeats. Canonical Length Sequences: The number of gap-occupying sequences that have the canonical length of Armadillo repeats.

Figure 2.10 shows a selection of MSAs from Table 2.2. Selected alignments will be focused on to describe the performance seen across all alignment methods. Although ClustalΩ did not perform well, the MSAs with GLProbs in Figure 2.10A, as well as ProbCons and MSAProbs (both not shown), have insertions between nearly all high-occupancy columns (39/41; 95%). Although the sequence logo (not shown) agrees with the conserved positions in the literature, these MSAs are not an accurate representation of this repeat family due to numerous insertions. Although having not nearly as many insertions, MSAs from MUSCLE
Development of RAM

2.3 Methods and Contents

(21/40; 53%) in Figure 2.10B, as well as MAFFT and ClustalΩ (both not shown), also have gaps within SSEs. The best alignment method is ClustalW, shown in Figure 2.10C, having only seven insertions across the whole alignment. Although ClustalW is a considerable improvement over the other alignment methods discussed so far, there are still insertions within all three helices of Armadillo repeats. Finally, in Figure 2.10D, AMPS stands out from all the other alignment methods as having insertions between the SSEs annotated in the template alignment. Although the other alignment methods can be further improved by adjusting the various gap penalties, amongst other parameters, the alignment is likely to require manual curation. In contrast, AMPS builds an alignment that requires no further editing or realignment.
Figure 2.10: Four MSAs built using different alignment methods. Methods include (A) GLProbs, (B) MUSCLE, (C) ClustalW, and (D) AMPS. The sequences represent equidistant members of an average distance tree, calculated in Jalview from the full alignment. All annotations above and below the MSAs refer to information from the full-length MSAs. Annotation above MSAs: The annotation ladder (black) numbers the alignment columns in
intervals of 5 amino acids. Low occupancy alignment columns, less than 50% residue occupancy, are hidden. Arrays of hidden columns are collapsed and depicted as a single blue line with an arrowhead above the alignment. Annotation below MSA: two annotation tracks below alignments that summarise the information for the full-length MSA; Occupancy relates to the number of amino acids occupying each alignment column; SHENKIN is $PR_{Shenkin}$, where lower scores indicate the most conserved positions. All alignments are coloured by $PR_{Shenkin}$ shaded with a white (unconserved) to a red (conserved) colour scale.
2. Development of RAM

2.3 Methods and Contents

Figure 2.10 (cont.)
2.3.5 Multiple sequence alignment analysis

Sequence analysis entails the application of different methods for scoring residue conservation (Valdar, 2002) to gain evolutionary insights into an MSA. However, repeats belonging to class three of solenoid domains have few positions that are required to be conserved for the structural integrity of the fold (Kajava, 2012). Therefore, using one or more conservation scores can provide a limited view of diverse protein families such as repeats. This section describes elaborations of in-depth sequence analyses that extend beyond traditional approaches for studying an MSA. Due to the diversity in repeat sequences, this section continues discussing how to carry out subfamily analysis and identify why canonical positions are unconserved across the whole alignment.

2.3.5.1 Identifying evolutionarily conserved residues

RAM supports 18 different amino acid scoring methods employed with Amino Acid Conservation (AACon) software (Golicz et al., 2019, MacGowan et al., 2020), and Valdar (2002) discusses most of them in this review. However, the primary scoring method implemented into RAM is the symbol frequency-based Shenkin score (Shenkin et al., 1991). $V_{\text{Shenkin}}$ is a divergence score defined as:

$$V_{\text{Shenkin}} = 2^S \times 6 \quad (2)$$

$$S = - \sum_{i}^{K} p_i \log_2 p_i \quad (3)$$

Shenkin et al. (1991) proposed equation 1, where $S$ is Shannon’s entropy (Equation 2), $K$ are all 20 amino acid types, and $i$ represents the amino acid making up $K$: $V_{\text{Shenkin}}$ scores between $6 \leq V_{\text{Shenkin}} \leq 120$. For the most conserved, with a single amino acid type, a frequency of 1 would give an entropy of 0. Thus, $V_{\text{Shenkin}} = 2^0 \times 6 = 6$ for a conserved position. Suppose all 20 amino acid types were present at an alignment column with a frequency of $1/20 = 0.05$. In that case, an entropy of $S \approx 4.32$ produces a maximum $V_{\text{Shenkin}} 2^{4.32} \times 6 \approx 120$ for an
unconserved or diverse position. $V_{\text{Shenkin}}$ is transformed twice to aid with interpretation when scoring an MSA. First, Javier normalised $V_{\text{Shenkin}}$ to 0–100, defined as:

$$N_{\text{Shenkin}} = \frac{(V_{\text{Shenkin}} - V_{\text{Shenkin}}_{\text{min}})}{(V_{\text{Shenkin}}_{\text{max}} - V_{\text{Shenkin}}_{\text{min}})}$$

Lastly, all $N_{\text{Shenkin}}$ values across a given alignment are converted to percentile rank (PR) to express all the values as $0 \leq \text{PR}_{\text{Shenkin}} \leq 100$. By converting $N_{\text{Shenkin}}$, RAM can assign a conservation threshold for consistency across various repeat families with different conservation profiles. PR$_{\text{Shenkin}}$ is particularly useful for the TPR family (Section 3.4.2.3), where the alignment is split into three separate alignments with a different number of sequences. The $N_{\text{Shenkin}}$ percentile rank is given by:

$$PR = \frac{CF - (0.5 \times F)}{N} \times 100$$

CF is the Cumulative frequency that determines the number of observations less than or equal to a particular score of interest, $F$ is the frequency for the score of interest, and $N$ represents the number of scores in the distribution. Now RAM can assign a default sequence conservation threshold ($\text{PR}_{\text{Shenkin}} < 25\%$) to all repeat families and identify the most conserved positions in an MSA.

### 2.3.5.2 Classifying conserved physicochemical properties

Repeats can accommodate multiple amino acids at certain positions. Despite some positions scoring low with PR$_{\text{Shenkin}}$, some positions often share similar physicochemical properties that a single conservation score does not capture. An alternative approach to complement standard scoring methods involves a quantitative summary of physicochemical properties at every position in the MSA. The amino acid property index, defined in Livingstone and Barton (1993), was used here to provide the properties describing each amino acid. For example, position 9, characterised as buried in Section 5.5.3.1, is classified as unconserved with a PR$_{\text{Shenkin}}$ of 71.4. By summing the physicochemical properties, the top three properties
in Table 2.3 are hydrophobic, aliphatic, and small. Classifying positions based on the most conserved set of properties adds an additional layer of information to each position.

Table 2.3: Top three most common physicochemical properties at each alignment position for the Armadillo repeat MSA in Figure 2.9. The physicochemical properties use the amino acid property index, as described by Livingstone and Barton (1993), from the Venn diagram defined by Taylor (1986). First, each amino acid's physicochemical properties were obtained in the MSA. The properties were then summed for each alignment column and sorted in descending order of frequency to show the top three most common properties.

<table>
<thead>
<tr>
<th>Alignment Column</th>
<th>Rank</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Polar</td>
<td>Small</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Polar</td>
<td>Small</td>
<td>Charged</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Polar</td>
<td>Small</td>
<td>Charged</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Polar</td>
<td>Small</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Charged</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Small</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Polar</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Polar</td>
<td>Small</td>
<td>Charged</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Small</td>
<td>Hydrophobic</td>
<td>Polar</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Small</td>
<td>Hydrophobic</td>
<td>Tiny</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Tiny</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Small</td>
<td>Polar</td>
<td>Proline</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Polar</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>Polar</td>
<td>Charged</td>
<td>Small</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Small</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>Polar</td>
<td>Small</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>Small</td>
<td>Polar</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>Polar</td>
<td>Small</td>
<td>Charged</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>Small</td>
<td>Polar</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>Polar</td>
<td>Charged</td>
<td>Small</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Charged</td>
</tr>
</tbody>
</table>
## 2. Development of RAM

### 2.3 Methods and Contents

<table>
<thead>
<tr>
<th></th>
<th>Polar</th>
<th>Small</th>
<th>Hydrophobic</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Tiny</td>
</tr>
<tr>
<td>33</td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Aliphatic</td>
</tr>
<tr>
<td>34</td>
<td>Hydrophobic</td>
<td>Polar</td>
<td>Small</td>
</tr>
<tr>
<td>35</td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Tiny</td>
</tr>
<tr>
<td>36</td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>37</td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Small</td>
</tr>
<tr>
<td>38</td>
<td>Polar</td>
<td>Small</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td>39</td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
</tr>
<tr>
<td>40</td>
<td>Small</td>
<td>Hydrophobic</td>
<td>Tiny</td>
</tr>
<tr>
<td>41</td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Small</td>
</tr>
</tbody>
</table>

Alignment Column: High-occupancy alignment column in the MSA of Figure 2.9. Rank: The top three most common physicochemical properties at each alignment position.

The advantages of using this approach would help summarise the features of specific positions that are not as conserved. In addition, for highly divergent sites involved with binding-specificity, this is useful to characterise the amino acids that tend to reside at those positions. The only caveat is that describing positions in this manner oversimplifies the characteristics of positions that may be understood better by doing subfamily analysis, as discussed in the next section.

#### 2.3.5.3 Subfamily analysis

While summarised physicochemical properties can characterise an MSA, a more sensible approach would be subfamily analysis. This entails splitting the alignment into sensible groups, described in Section 2.3.6. It is beneficial to categorise protein sequences into individual subfamilies based on structure, function, sequence similarity or other criteria (Livingstone and Barton, 1993). Furthermore, alignment methods make pairwise comparisons when building MSAs, and a tree-based alignment can be used (Barton, 1994, Notredame et al., 2000, Sievers et al., 2011, Larkin et al., 2007).
2.3.5.3.1 Alignment dissection by repeat architecture

The capping repeats of linear solenoid domains have different solvating properties from those of middle repeats since they only fold with an adjacent repeat. Splitting the alignment by start, middle, and end repeats brings out the differences in the amino acid conservation profiles. Figure 2.11 shows three TPR MSAs representing the global repeat architecture types: start, middle, and end. The differences between start and end TPR MSA conservation profiles (PRShenkin), assembled in the grouped bar plot of Figure 2.11D, are ranked in descending order in Table 2.4. The largest changes in conservation are two pairs of positions: 3 and 10 are conserved in middle and end TPRs, with a PRShenkin of 26.5% (middle) and 35.3% (end), and 32.4% (middle) and 23.5% (end), respectively; 12 and 17 are only conserved in start TPRs, with a PRShenkin of 50.0% and 20.6%, respectively.

On studying these pairs of residues, positions 3 and 10 have sidechains that face the previous TPR in the N-terminus direction. In contrast, positions 12 and 17 face the next TPR in the C-terminus direction. The observed trends in the conservation profiles of MSAs in Figure 2.11A-B are structurally validated in panel E. Furthermore, this subfamily analysis identifies residues likely to be essential for the fold of TPR-Domains, and it can be extended to other linear solenoid domains. Subfamily analysis can go further by clustering together repeats that share sequence homology. The following section discusses how this approach can group repeats that may be functionally or evolutionarily related.
Figure 2.11: Extracts of each TPR alignment in Homo sapiens split by TPR position. The (A) start, (B) middle, and (C) end alignments have 17 sequences. All alignments are coloured by PRShenkin. The sequences represent equidistant members of an average distance tree, calculated
in Jalview from the full alignment. Annotation above the alignment: The annotation ladder (black) numbers the alignment columns in intervals of 5 amino acids. As discussed in Section 3.2, the arrowheads with numbers (green) denote the conserved residues highlighted in the literature. Based on the PRShenkin histogram, the MSA is coloured with a white (unconserved) to a red (conserved) colour scale. The secondary structure is highlighted as individual boxes labelled as Helix A, the first linker (L1), Helix B, and the second linker (L2). Low occupancy alignment columns, less than 50% residue occupancy, are hidden. Arrays of hidden columns are collapsed and depicted as a single blue line with an arrowhead above the alignment. The representation and graphical output were generated in the Jalview. (D) Grouped bar plot of the PRShenkin across all three alignments: start MSA in blue, middle in orange, and end in green. (E) Adjacent TPRs highlighting the four conserved positions with respect to repeat architecture.

Figure 2.11
(cont.)
Table 2.4: Differences in $\text{PR}_\text{Shenkin}$ between start and middle TPRs.

<table>
<thead>
<tr>
<th>Alignment Column</th>
<th>$\Delta \text{PR}_\text{Shenkin}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.5</td>
</tr>
<tr>
<td>2</td>
<td>26.5</td>
</tr>
<tr>
<td>3</td>
<td>58.8</td>
</tr>
<tr>
<td>4</td>
<td>2.9</td>
</tr>
<tr>
<td>5</td>
<td>8.8</td>
</tr>
<tr>
<td>6</td>
<td>11.8</td>
</tr>
<tr>
<td>7</td>
<td>17.6</td>
</tr>
<tr>
<td>8</td>
<td>0.0</td>
</tr>
<tr>
<td>9</td>
<td>23.5</td>
</tr>
<tr>
<td>10</td>
<td>64.7</td>
</tr>
<tr>
<td>11</td>
<td>5.9</td>
</tr>
<tr>
<td>12</td>
<td>38.2</td>
</tr>
<tr>
<td>13</td>
<td>5.9</td>
</tr>
<tr>
<td>14</td>
<td>17.6</td>
</tr>
<tr>
<td>15</td>
<td>5.9</td>
</tr>
<tr>
<td>16</td>
<td>23.5</td>
</tr>
<tr>
<td>17</td>
<td>44.1</td>
</tr>
<tr>
<td>18</td>
<td>20.6</td>
</tr>
<tr>
<td>19</td>
<td>14.7</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
</tr>
<tr>
<td>21</td>
<td>5.9</td>
</tr>
<tr>
<td>22</td>
<td>0.0</td>
</tr>
<tr>
<td>23</td>
<td>8.8</td>
</tr>
<tr>
<td>24</td>
<td>8.8</td>
</tr>
<tr>
<td>25</td>
<td>0.0</td>
</tr>
<tr>
<td>26</td>
<td>11.8</td>
</tr>
<tr>
<td>27</td>
<td>5.9</td>
</tr>
<tr>
<td>28</td>
<td>11.8</td>
</tr>
<tr>
<td>29</td>
<td>2.9</td>
</tr>
<tr>
<td>30</td>
<td>26.5</td>
</tr>
<tr>
<td>31</td>
<td>5.9</td>
</tr>
<tr>
<td>32</td>
<td>2.9</td>
</tr>
<tr>
<td>33</td>
<td>0.0</td>
</tr>
<tr>
<td>34</td>
<td>17.6</td>
</tr>
</tbody>
</table>
Alignment Column: High-occupancy alignment columns in the full-length MSA. \(\Delta \text{PR}_{\text{Shenkin}}\): Absolute differences in PR\(_{\text{Shenkin}}\) between repeat architectures.

2.3.5.3.2 Clustering and cutting the tree by a dendrogram

A straightforward approach is to do pairwise comparisons and cluster sequences from those scores, as described in Section 2.3.6. This strategy can cluster repeats that share a common property without a priori knowledge of a repeat family. For example, Figure 2.12 visualises a cropped average distance tree for the TPR MSA (from Section 3.4.2.2). The tree was cut at a particular position, highlighting three large and distinct subfamilies in Figure 2.12 that reflect the function, evolution, and PRA of 3-TPR-Domains. All the proteins, across all three subfamilies of MSAs, bind to Heat-shock proteins 70 and 90 (Scheufler et al., 2000). Each subfamily comprises the individual TPRs of a 3-TPR-Domain, highlighting how clustering can emphasize similar repeat architecture. Furthermore, positions unconserved in the alignment become conserved within the subfamilies, thus highlighting the functionally relevant residues important for binding to the MEEVD motif of HSP70/90 proteins.
Figure 2.12: Three TPR subfamily MSAs with each MSA making up a TPR in a 3-TPR-Domain. MSAs are coloured with ClustalX colour scheme.
Figure 2.11 (cont.)
2.3.6 maxHMM

Evaluating the coverage of annotations from a dataset and determining how comprehensive they are will establish whether the repeat family of interest is sufficient for downstream analyses. Often individual databases can have high coverage, having identified many repeat-containing proteins. However, many proteins have annotations that are not comprehensive and are disjointed in the primary sequence, as in Figure 2.2B. The resulting outcome is a fragmented dataset with poorly annotated Repeat-Domains. The assessment of the dataset’s quality (Section 2.3.1), studying the PRA (Section 2.3.2), and understanding the

![Figure 2.11](cont.)
underlying sequence (Section 2.3.5), will determine whether the existing set of repeat annotations need to be reannotated and missing repeats filled in.

Using a general HMM to summarise the repeat family can incorporate too much sequence diversity, and the sequence profile struggles to recapture the training dataset confidently. Suppose the non-redundant set of annotations used is not comprehensive: In that case, maxHMM is a pipeline that builds a multi-level hidden Markov model library trained on sequences derived from an MSA. The approach was adopted from the method described for protein kinases by Miranda-Saavedra and Barton (2007) to overcome the limitations of a single HMM profile. MaxHMM can be used to reannotate existing repeats, identify previously undetected repeat annotations, and characterise them. MaxHMM was demonstrated using the TPR family to fill in the missing gaps.

Conceptually summarised in Figure 2.13, the method loops around five steps: (1) input an MSA by either building one (Section 2.3.3) or using an existing one; (2) calculate Z-scores from pairwise alignments of sequences; (3) cluster sequences by Z-scores and build a dendrogram; (4) cut the dendrogram to generate subfamilies; (5) build HMM-profiles from the resulting subfamilies with HMMer (Eddy, 2011); (6) search database of proteins with the library of HMM-profiles with HMMer; (1’) chain new sequences to the main MSA and restart the loop; (7) The loop ends when no new sequences are detected by maxHMM.
2.3 Methods and Contents

2.3.6.1 Generating and searching with a multi-level HMM library

In step 2 of Figure 2.13, the AMPS suite of programs performs pairwise comparisons of all sequences in the MSA, from step 3 in Figure 2.13. First, all unique pairs of sequences in the alignment are optimally aligned with the Needleman–Wunsch algorithm (Barton and Sternberg, 1987, Needleman and Wunsch, 1970), as programmed in the AMPS package using
amino acid substitution scores from BLOSUM62 (BLOcks SUbstitution Matrix) substitution matrix. Furthermore, a constant value and fixed-length gap-penalties are two parameters that can be assigned to modulate the alignment step (Barton and Sternberg, 1990). Next, the sequences are shuffled and compared again for a set number of times. The purpose of this is to determine the expected distribution of scores had the pair of sequences been unrelated, with an identical length and amino acid composition, as in the pair of unshuffled sequences. Lastly, AMPS calculates the Z-scores, which is the number of standard deviations from the mean of scores after aligning a set number of randomised pairs of sequences. The histogram in Figure 2.14 shows the distribution of all Z-scores for the TPR MSA (Chapter 3) with 100 randomised pairs of sequences using a BLOSUM62 matrix.

![Z-score histogram](image)

Figure 2.14: Z-score histogram for pairwise sequence comparisons of the TPR MSA. Histogram for the output of the TPR MSA of step 2 in maxHMM (Figure 2.13). Parameters used to generate these data include: 100 randomised pairs for each pairwise comparison, a constant value of 0, a gap penalty of 10, BLOSUM62 matrix.

In step 3 of Figure 2.13, the OC program builds a dendrogram from the pairwise data by applying complete linkage clustering analysis to the Z-scores (Barton, 2002). MaxHMM can then select clusters according to one or multiple Z-score cut-offs.
Sequences within each cluster were dissected from the original MSA into a separate cluster MSA. Next, the hmmbuild program from the HMMER suite of programs (HMMER version 3.1b2; Eddy, 1995) builds an HMM profile from each cluster MSA to produce a library of HMM profiles. The hmmbuild program uses an extremely low inclusion threshold (-10), which is used to determine what matches should be considered ‘true’, regardless of their sequence bit score. Finally, the hmmsearch program searches full-length protein sequences from the UniProtKB database with each HMM profile from the library.

2.3.6.2 Processing hits
MaxHMM generates all the hmmsearch hits, collates and post-processes them in the workflow displayed in Figure 2.15. MaxHMM processes raw hmmsearch hits both automatically and manually. The automatic stage includes four steps: (1) post-process of all hits, termed raw hits, from the standard hmmsearch output files; (2) extending each raw hit to the length of the HMM-profile that it was detected by, termed extended hits, and remove raw hits that extend beyond the N- and C- termini of the protein; (3) maxHMM groups all extended hits into sets with JI (Section 2.3.1.2) and the highest scoring extended hit, termed match, within each set are selected while maxHMM discarded the others; (4) automatically labelling matches based on whether they overlap with the existing repeat annotations that they were trained on for the multi-level HMM library, or annotations that represent the repeat family of interest across all Biological Entities from InterPro.

The manual stage includes one step, where low scoring matches across all proteins are manually assessed for additional supporting evidence. The first step is to refer to the labels (step 4) and understand how the matches compare with other related annotations across all Biological Entities. Next, low scoring matches would be further substantiated by studying the PRA to identify gaps resembling the length of a canonical repeat. Other supporting elements involve referring to solved structures from the PDBe or predicted structure models, for
example, PHYRE2 (Kelley et al., 2015). Finally, studying the underlying sequence of low-scoring matches.

Figure 2.15: Processing hits from the hmmsearch tool. The workflow by which maxHMM processes hits automatically (steps 1-4) and manually (step 5) was demonstrated on a single and hypothetical protein. (1) post-process of hmmsearch output. The empty white box is the HMM profile, which encases the raw hit with hmmsearch. (2) Extending each hit to the length of the HMM profile. (3) Removing low-scoring hits and retaining matches. (4) Labelling matches. (5) Manual curation of matches with support from PRA analysis, structure comparison, and underlying amino acid sequence. The symbol ‘X’ above annotations means removing an annotation. In step 5, capitalised letters H (blue), D (green), and R (orange) refer to the biological entity Homologous Superfamily, Domain, Repeat, respectively. Multiple rows within each biological entity represent the contributing member databases, respective to the
colour of the biological entity in the figure. The dark grey annotations within each biological entity are the non-redundant InterPro annotations, summarising the annotations from member databases.

2.3.6.3 Automatic stage

The first step, in Figure 2.15, parses the hmmsearch output from each HMM profile in the library. MaxHMM then groups all raw hits in the correct format for downstream analyses. The final technical process extends the terminal positions of all raw hits corresponding to the HMM profile.

When selecting for high-scoring matches, maxHMM groups matches that annotate a similar region of a protein, using the method in Section 2.3.1.2 (Figure 2.3). Within each group, maxHMM retains the match with the highest bit score and discards the other matches. This process loops using descending jT values from 0.9 to 0.2 that decrease by 0.1 with every iteration. This strategy incrementally groups near-identical matches to select the highest scoring one within each group. The purpose of applying such a method is to avoid incorrect and low-scoring hits from chaining together high-scoring extended hits that correctly recognise an actual repeat. As a result, the process would have accidentally removed genuine matches if not for this approach.

MaxHMM is provided with three datasets that make up the entire InterPro dataset, for either single or multiple species: annotations that represent real repeats, other annotations on different Biological Entities that also recognise regions that have the repeat of interest, and annotations that do not label this repeat family.

The last phase of the automatic stage (step 4) labels all matches based on multiple aspects. The first part categorises maxHMM matches into three initial labels, termed Seed Label: Seed repeat refers to whether a match overlaps (jT>0.2) with an annotation that was trained on by hmmbuild; Non-repeat (jT>2) refers to an annotation that does not represent this repeat family; other refers to neither of the other two.
Supplementing the three initial labels, maxHMM adds additional labels to the matches based on their overlap (\(jT>0\)) with annotations, representing the family of interest, singled out from the true negative dataset: Repeat Region is a broad classification referring to annotations from other Biological Entities that correctly label the repeat family of interest, for example, the green annotations in the Domain biological entity of Figure 2.15 (step 5); Repeat-like Helical Domain annotates much larger regions predicted to have multiple repeating structural motifs, for example, the blue annotations in the Homologous Superfamily biological entity of Figure 2.15 (step 5); Finally, any remaining matches with no label will be labelled Unassigned.

Lastly, the final labels are the standard PRA analysis from Section 2.3.2. Since repeats are adjacent to form Repeat-Domains, identifying matches that are adjacent to other ones is a helpful metric to measure. In addition, this analysis helps discard low-scoring repeats in isolation, which are likely contributing to a much larger domain that is not a Repeat-Domain.

MaxHMM identifies the most optimal bit score threshold by calculating Matthews correlation coefficient (MCC; Matthews, 1975) that best classifies actual repeats from annotations that are not repeats. In this case, true positives identify seed repeats that meet the threshold, whereas true negatives are non-repeats below the threshold. False negatives are seed repeats that did not meet the threshold, and false positives are non-repeats scoring above the threshold. To better understand how some matches are falsely classified and what the other matches represent, the manual stage of maxHMM attempts to delineate them.

2.3.6.4 Manual stage

False negatives can often be matches that represent the termini of Repeat-Domains and naturally score low. In contrast, false positives are often incorrectly classified so. Often, false positives tend to pair with labels from step 4 that would indicate otherwise. For example, observed with the TPR dataset in Section 3.3.3.2, false-positive matches would be labelled Repeat Region and Repeat-like Helical Domain, overlapping with an annotation from the true negative dataset. These annotations belong to the TPR Clan (CL0020).
The TPR Clan is an extensive collection of Pfam entries that annotate the TPR, TPR-like, Armadillo, HEAT repeat families, amongst other minor all-helical repeat families. With a collection of 149 related Pfam entries (Pfam release version 33.0), profile–profile comparisons indicate that they significantly share a similar sequence and have analogous all-helical structural folds and related functions. Therefore, false positive matches with supporting repeat labels are reviewed manually to ensure they are included in the final dataset.

Other matches are of interest here since they are often putative repeat annotations. Other matches that score below the threshold require additional support to include them in the final dataset. These include studying the labels assigned to them. The strongest candidates for inclusion are likely to have both Repeat Region and Repeat-like Helical Domain adjacent to other matches or seed repeats. Additionally, other matches can be referenced to a solved or modelled structure to support them. Weak candidates only overlap with Repeat-like Helical Domain annotations, and these candidates tend to have no structures and are often resolved with PRA analysis and studying the underlying sequence.

2.3.7 Human population variant analysis

RAM implements a novel approach to exploring human population variant data, from the Genome Aggregation Database (gnomAD; Karczewski et al., 2020b), in human repeat sequences. This work extends the key concepts explored in a previous study across all Pfam families (MacGowan et al., 2017), thus learning more about each repeat family with human population variant analysis.

2.3.7.1 Mapping variants and measuring human population constraint

VarAlign combines tools, pipelines, and resources into a single pipeline. The primary function of VarAlign is to retrieve the human population variant data and map it to an MSA (MacGowan et al., 2017). Figure 2.16 shows an overview of all the main components of the pipeline.
Figure 2.16: Overview of the VarAlign pipeline. VarAlign retrieves variants from gnomAD and maps them to human sequences in an MSA. Next, VarAlign encapsulates the ProIntVar pipeline and employs it for structural analysis. ProIntVar retrieves all structures for sequences in the MSA that have at least one structure solved. ProIntVar further employs Arpeggio to calculate interatomic interactions and run DSSP to obtain the secondary structure definitions, and residue surface accessibility (RSA) values. Finally, VarAlign maps the structural data to the MSA. Figure 5 from Utgés et al. (2021).

VarAlign maps missense and synonymous variants from gnomAD (version 2.0.1) to human sequences in an MSA, as demonstrated in Figure 2.17. VarAlign aggregates variants on a column basis and sums them; termed variant column totals.
2. Development of RAM

2.3 Methods and Contents

Figure 2.17: Schematic representation of mapping and aggregating variants. VarAlign maps to a hypothetical MSA with three sequences. Sequences are depicted as a thin black line, where white and empty circles represent variants mapped to the sequence. Underneath the alignment are the variant column totals, labelled ‘T’, that sum variants at each alignment column.

Next, VarAlign measures the proportion of variants at each column relative to other positions in an MSA. It does so by calculating a missense enrichment score (MES), defined as:

\[
MES = \ln \left( \frac{Var_i/Var_t}{Occ_i/Occ_t} \right)
\] (6)

MES is expressed as the natural logarithm (log) of the odds ratio (OR; Szumilas, 2010) of all variant column totals to associate and compare with the rest of the MSA. Using the example in Table 2.5, the formula for MES entails the division between variant column total of one column with the sum of the other variant column totals. The formula then divides this value by dividing column residue occupancy of the same single column with the sum of the other residue column occupancies. The formula is applied to all columns in an MSA. RAM finally classifies the positions as enriched, depleted, or neutral for human missense variants relative to other positions in the MSA.

Table 2.5: Odds ratio table for variants in an MSA. RAM calculates the OR for the number of variants in a single alignment column relative to the rest of the hypothetical MSA. Using Equation 6, the OR 5/100 divided by 50/500 equals 0.5. The logOR is ≈ -0.69.
Occancy: The number of residues that make up an alignment column. Variants: Number of variants. Column: A single column in an MSA. Column: The rest of the columns in the MSA that are not Column.

RAM also retrieves variants from ClinVar data in the same manner as for the gnomAD dataset. This was applied to the TPR and Ankyrin repeat annotations from human sequences in Chapters 3 and 4. Version 1.53 of the ClinVar dataset was used in this analysis.

2.3.7.2 Classifying positions

Alignment columns of an MSA are classified by their constraint in the human population from the distribution of missense variants using MES and sequence conservation from conventional sequence analysis using $PR_{Shenkin}$.

The missense enrichment plot for the Armadillo repeat family, in Figure 2.18, shows the MES (y-axis) is plotted against $PR_{Shenkin}$(x-axis). A regression line was modelled to this data to understand the trend. RAM classifies all alignment columns of the Armadillo repeats into four categories according to MES around the regression line and $PR_{Shenkin}$. Additionally, a fifth, unclassified, category collectively groups positions that do not meet the thresholds. Columns with $0 \leq PR_{Shenkin} \leq 25$ and residual $< 0$ are classified as conserved and missense depleted (CMD), whereas $0 \leq PR_{Shenkin} \leq 25$ and residual $> 0$ are classified as conserved and missense enriched (CME). Other classifications included columns with $75 \leq PR_{Shenkin} \leq 100$ that are unconserved and missense depleted (UMD) with residual $< 0$ and missense enriched (UME) with residual $> 0$. 

<table>
<thead>
<tr>
<th></th>
<th>Column</th>
<th>Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variants</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>Occupancy</td>
<td>50</td>
<td>500</td>
</tr>
</tbody>
</table>
Supplementing the classification of sites, RAM can study the variant properties of positions of interest. Inspired by the BLOSUM matrix (Henikoff and Henikoff, 1992), which describes the frequency of amino acid mutations over an evolutionary time, RAM can generate a substitution matrix of all the variants present at any position to understand the properties and number of variants. Figure 2.19 shows substitution matrices for two positions in Figure 2.18, demonstrating how position 35 has many variants ranging from alanine to valine or threonine. In contrast, position 30 is dominated by amino acid changes from charged residues such as arginine and glutamate, to cysteine, tryptophan, histidine, glutamine, and lysine.

A substitution matrix is grouped by amino acid physicochemical properties from the amino acid property index, defined in Livingstone and Barton (1993). Substitution matrices can provide an insight into how positions continue to evolve. For example, the top three most frequently observed amino acids at position 35 in Armadillo repeats (Table 5.1) are Alanine...
(38.2%), Threonine (14.4%), and Valine (11.3%). Position 35 is dominated by variants that mutate from alanine to the next two topmost frequent amino acids, possibly indicating that this site continues to evolve in a pool of amino acids with favourable physicochemical properties.

Figure 2.19: Substitution matrices for the Armadillo MSA. These characterise the amino acid properties of the missense variants at positions 35 (A) and 30 (B). Black lines split groups of amino acids that represent the ones from the amino acid property index, defined in Livingstone and Barton (1993).

Structural analysis

To explore and validate positions of interest, RAM carries out a comprehensive proteome-wide structural analysis by employing VarAlign. For example, in Figure 2.16, ProIntVar retrieves structures from the PDBe, and VarAlign then organises the structural properties of an alignment for downstream analyses.

For a given alignment, ProIntVar maps residues to PDB structures with SIFTS (Dana et al., 2019) for sequences with at least one solved structure. Arpeggio calculates all the interatomic interactions for each residue observed in one or more mapped structures (Jubb et al., 2017). Arpeggio considers two atoms to interact if they are within 5 Å of each other. Furthermore, ProIntVar runs DSSP (Kabsch and Sander, 1983) on all structures to determine the SSE and the residue relative surface accessibility (RSA). ProIntVar finally assembles all
2. Development of RAM

2.3 Methods and Contents

structural attributes from the PDBe, Arpeggio, and DSSP with respect to alignment column positions.

2.3.8.1 Contact maps

Positions with the largest number of residue pair contacts are probably the most important for maintaining the fold of a structure. Accordingly, RAM generates contact maps for each repeat for protein superimposition and characterises the similarities across all repeats.

RAM identifies each residue pair contact, supported by at least one interatomic interaction, for every repeat observed in a structure. The same principle applies to repeats with substrate interactions in the next section. Multiple structures have been solved for proteins; thus, to integrate this depth of information without overcounting residue pair contacts, RAM addresses the redundancy of the structural data in Figure 2.20. The redundancy observed here includes multiple structures solved for a single protein; and individual structures can have more than one chain. For the example in Figure 2.20A, this Armadillo repeat (194–234aa) is observed in five structures across three PDBe entries, where the third one (4R11) has three chains. RAM first calculates the five contact matrices (Figure 2.20C), combine them (Figure 2.20D) then collapse all the observations to a single observation (Figure 2.20E), termed Evidence of Contact (EoC). The EoC was defined in MacGowan et al. (2017) and implemented into RAM as a Pandas operation (Section 2.4) on the structure table from VarAlign and ProtIntVar. RAM iterates this process for all repeats in the MSA and combines these contact matrices (Figure 2.20F) to create a single contact matrix that summarises the EoCs across the Armadillo repeat family. Individual cells in a contact matrix represent a residue pair contact. The shade of the cell depicts the number of EoCs. The colour gradient to the left of the contact matrix in Figure 2.20G visualises this information.
2. Development of RAM

2.3 Methods and Contents

Figure 2.20: Building contact maps. RAM incorporates the full depth of structural data and tackles redundancy: (A) identifying all structures for the repeat (193–234aa) in Beta-catenin-like protein hmp-2 (O44326; HMP2_CAEEL); (B) retrieving the structural data; (C) calculating all contact matrices; (D) sum them; (E) collapse all observed contacts into a single
observed EoC information; (F) loop through all repeats in the alignment with at least one structure. This example was demonstrated with structural data from the Armadillo repeat family in Chapter 5.

The data supporting the contact map was split into intra- and inter-repeat contacts to dissect the network of interactions in a Repeat-Domain. Intra-repeat contacts represent residue pair interactions within a repeat, whereas inter-repeats are between repeats. Figure 2.20F combines intra- (top right half) and inter- (bottom left half) Armadillo contacts.

In the interest of studying the most frequently observed residue pair contacts, RAM applies a default EoC threshold at 50%. RAM identifies residue pair contacts that meet the EoC threshold and measures the occurrence of each alignment position interacting with another alignment position, termed contact number (CN). RAM then sets a threshold that is the average of all contact numbers and positions that meet the contact number threshold are interpreted as likely to be important for the Repeat-Domain fold.

2.3.8.2 Studying sites important for the structural motif

RAM identifies positions with a high contact number to delineate the most important sites in a repeat. Firstly, the long-distance interactions are considered, therefore RAM filters out all contact pairs under five amino acid positions apart. These are discarded because they make up the local backbone interactions that run the entire length of a repeat. These local interactions make up the long white strip in Figure 2.20F that runs from the top left corner to the bottom right corner of the contact matrix.

2.3.8.3 Direct coupling of adjacent repeats

Positions involved in inter-repeat interactions with the highest CNs can be further characterised by classifying the direction in which they interact. This is to further probe the different roles of each position in the fold of a Repeat-Domain. For example, interactions in the N- to C- terminus direction are termed forward interactions, whereas C- to N- terminus direction are termed backward interactions.
2. Development of RAM

2.3 Methods and Contents

2.3.8.4 Characterising different substrate-binding modes of Repeat-Domains

RAM uses the same strategy for substrate interaction analysis here too. EoCs between a single repeat and all its substrates observed in a structure, termed protein-protein interactions (PPI), are summarised at each alignment column.

When compiling PPI data, it is assumed that each repeat that can bind to more than one substrate will share the same mode of binding. Thus, a similar set of positions would bind with different substrates. Therefore, a PPI comprises a single count across all substrate interactions, for each repeat, at each alignment position.

2.3.8.5 Aggregated structural properties

This section discusses how the structural information from DSSP is used to describe repeat families further. Javier used the output from DSSP to characterise the structural properties of the Ankyrin repeat family; this analysis was integrated into RAM as a permanent component of structural analysis and was later applied to the TPR and Armadillo repeat family.

First, DSSP calculates SSE assignment and provides the RSA of each residue for a given protein structure. Then, using the same column-basis strategy, this structural data is aggregated and summarised, in Figure 2.21, to quantitatively assign the SSEs and characterise the RSA properties of each position.
Figure 2.21: Secondary structure assignment and RSA classifications for the TPR family. The bar plots summarise the structural data from the TPR MSA in Section 3.4.2.1. (A) Median residue surface accessibility for each alignment position. The different bars represent the proportion of residues with structural coverage that fall into the different RSA classifications according to the specified thresholds: surface (RSA > 25), partially exposed (5 < RSA < 25), and core (RSA < 5). (B) The proportion of structures that have SSE as defined by DSSP, and 0.5 is the threshold that assigns what the most common SSE is present at each position.

2.3.8.5.1 RSA analysis

RAM determines the RSA and plots the median RSA for the TPR MSA (Section 3.4.2.1) in Figure 2.21A. Originally outlined in MacGowan et al. (2017), the DSSP solvent accessibility calculation of the RSA is described in Tien et al. (2013). From the median RSA, RAM classified each position as surface (RSA > 25), partially exposed (5 < RSA < 25), or core (RSA < 5).
2.3.8.5.2 SSE classification

DSSP assigned the SSE classifications to each position in the structure, and RAM summarised it in Figure 2.21B. The classification was assigned for positions that satisfy the 0.5 proportion of residues presenting the SSE. For TPRs in Figure 2.21B, the structural motif is described as the first helix in positions 1–13, followed by a three amino acid long loop (14–16), a second helix between 17–30, and the final loop (31–4).

2.4 Implementation and data analysis

The RAM computational framework was developed in Python (version 3.6.10) and related libraries. Each core module of RAM, such as PRA analysis or structural analysis, was written and encapsulated as individual or a collection of Jupyter Notebooks (Kluyver et al., 2016).

Protein annotations were studied by viewing individual entries via a web interface of the InterPro database (Mitchell et al., 2019). Furthermore, protein structures and their information were viewed on the webpage of the PDBe (Velankar et al., 2010). All dataframes, such as InterPro, UniProtKB/Swiss-Prot, and maxHMM output annotations, were processed with Pandas library (McKinney, 2011). The PRA analysis is solely analysed with Pandas for a given set of repeat annotations.

Statistical analyses were performed with SciPy (Virtanen et al., 2020) and Scikit-learn (Pedregosa et al., 2011) libraries. OR and Fisher’s exact $p$ values were calculated with the fisher_exact function from the stats library in SciPy. Linear regression models were calculated with the LinearRegression function from the linear_model library in scikit-learn. Mathematical functions were employed via the built-in module in Python for simple arithmetic tasks.

Plotting and visualisation were performed with Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) libraries. All structures were generated with UCSF Chimera (Pettersen et al., 2004). Figures were organised and assembled in Adobe Illustrator (Inc., 2019).
MSAs were handled with the biopython library (Cock et al., 2009). All alignments were finalised and stored in the Stockholm format, or in FASTA or BLC format where appropriate. MSAs were viewed and analysed, and dendograms were generated and studied in Jalview benchwork (Waterhouse et al., 2009). Alignment methods were either employed via AACon in Jalview or executed on the command line where appropriate.

2.5 Conclusions

RAM is the main computational framework developed in this PhD project. It integrates all the steps, tools, and methods to explore human population variation data in protein repeat families and support it with a comprehensive structural analysis. This project focused on designing RAM to address inconsistencies in the literature and provide a workflow compatible with all protein repeat families. At the same time, RAM introduces the novelty of integrating human population variation data and organising many structures to understand the function and evolution of repeats.

RAM was discussed with results borrowed from the results chapters of individual repeat families subsequently studied with RAM. Not all elements of RAM were applicable to all three repeat families. When studying the individual repeat families, each one proved to be a unique challenge that raised new components to implement into RAM.

The key features discussed in the RAM computational framework are as follows:

- RAM provides a framework for modular analysis of PRA analysis, genetic variation data, and structural data
- RAM provides clear guidelines and suggestions for studying the literature, selecting appropriate databases for repeat annotations, analysing PRA, building MSAs, scoring and analysing whole and subfamily MSAs, reannotating repeat annotation datasets, interpreting genetic variation, and structural analysis
• RAM implements the redefined and universal repeat nomenclature for novel PRA analysis, that can be extended to other repeat families
• Building and optimising MSAs with different methods to incorporate repeats with indels
• Scoring, assessing, and comparing MSAs from different alignment methods
• Applying a multi-level HMM library approach to discover new repeats and correct for fragmented annotation datasets
• Implements novel human population variation data to identify positions under selective pressure in the human population
• Lastly, it integrates a comprehensive and quantitative structural analysis that brings together an extensive collection of protein structures, and introduces a novel analysis to dissect the structural and functional roles of positions
3.1 Preface

TPR family established the core elements of RAM. TPRs initiated the development of the PRA analysis in RAM. It started as a simple question about how TPRs organised themselves in primary sequence from the review by D'Andrea and Regan (2003). PRA and nomenclature answered this question in much more depth and is now a mainstay in RAM to study PRA.

The next aspect of RAM development involved summarising physicochemical properties. In comparison to other repeat families, TPRs are the most unconserved repeat family in sequence (Figure 1.3). This stage aggregates amino acid properties in the MSA and quantitatively measures this information, which supplements standard methods for scoring residue conservation.

Difficulties in annotating TPRs set off the work with maxHMM to detect new TPRs and characterise them in terms of repeat architecture and sequence homology. MaxHMM helped appreciate the evolution of all repeat families that comprise the TPR superfamily.
Concatenating structural data across all repeats created the key structural analysis of RAM. Building off from MacGowan et al. (2017) who aggregated structural features across equivalent positions of MSAs, this analysis is the core analysis implemented with repeats. Since repeats present many more copies within a single species, studying the most frequent interactions across all structures are insightful. Thus, integrating high depth of structural data introduced something to the field of repeats that has not been previously attempted. This is the first attempt in the repeats community to characterise the structural roles of each position in terms of intra- (within) and inter- (between) repeat interactions.

3.2 Introduction

When Hirano et al. (1990) first discovered TPRs in the cell cycle protein CDC23, Sikorski et al. (1990) also found similar strings of sequences that resembled TPRs in three proteins involved with mitosis and two involved with RNA synthesis. Since then, they have been found in many proteins with different functions found throughout evolution (Jernigan and Bordenstein, 2015, D'Andrea and Regan, 2003, Blatch and Lässle, 1999, Kamel et al., 2021). Examples include the RNA polymerase-associated protein CTR9 homolog protein that makes up the PAF1 complex involved with transcription elongation (Deng et al., 2018); Apc3, Apc5, Apc6, Apc7, and Apc8 organise the anaphase promoting complex (APC) for cell cycle progression (Alfieri et al., 2017); O-GlcNAc transferase (OGT) protein involved with neurodevelopment (Howerton et al., 2013); Translocase of outer membrane 70 (TOM70) and Peroxisomal targeting signal 1 receptor (PTS1R) are involved with mitochondrial (Kreimendahl and Rassow, 2020) and peroxisomal (Gatto et al., 2000) protein transport, respectively.

Due to the amphiphilic properties of the helices, TPRs were originally predicted as an \( \alpha \)-helical structure proposed to fold into a coiled-coil or a helical-bundle structure with a ‘knob and hole’ packing (Hirano et al., 1990). The first structure of protein phosphatase 5 (PP5) with
TPRs, solved by Das et al. (1998), displayed a novel helical array fold at the time. Figure 3.1A shows canonical TPRs are 34 amino acids long that fold into a conserved anti-parallel helix-turn-helix structural motif with a packing angle of 24°, termed Helix A and B (Figure 5.1B) (Groves and Barford, 1999, Das et al., 1998). Unlike other helical repeat families with a hydrophobic core that extends across adjacent repeats, Figure 3.1B and D show that TPR-Domains are instead defined by a pattern of small and large hydrophobic residues that interlock in the helical-bundle structure. As a result, the topology of helices across adjacent TPRs fold in a corrugated arrangement that differ from Ankyrin, Armadillo, and HEAT repeats (Groves and Barford, 1999, Das et al., 1998). In Figure 3.1B, TPRs have a right-handed fold, where adjacent TPRs stacked together in a parallel arrangement to form a right-handed superhelical domain structure with a small curvature (Figure 3.1C).

Figure 3.1A shows TPRs are characterised by a degenerate consensus sequence. Despite its few conserved positions, the names “tetra” and “trico” refer to the four conserved sites on Helix A and three on Helix B, respectively – thus giving rise to its name (Sikorski et al., 1990). The tiny hydrophobic residues dominate TPRs at positions 8 (glycine and alanine) and 20 and 27 for alanine. These positions are vital for permitting helices within TPRs to fold intimately (Lamb et al., 1995, Blatch and Lässle, 1999, D’Andrea and Regan, 2003). Other positions also favour residues with specific physicochemical properties. For example, larger residues with aliphatic or aromatic sidechains tend to occupy the hydrophobic pockets between the helices of adjacent TPRs. Other positions in Figure 3.1A favour helix-breaking amino acids, where 15 favours glycine residues to allow for the tight turn between Helix A and B and punctuated by proline-induced turns at position 32 that defines the boundaries of a TPR.

Collectively, TPRs are organised in 3-16 tandem or dispersed arrangements with multiple combinations (Blatch and Lässle, 1999, D’Andrea and Regan, 2003). Although Blatch and Lässle (1999) began qualitatively assessing the PRA of TPR-Domains as new TPR-
containing proteins were being discovered, it was not until D'Andrea and Regan (2003) that they quantitatively measured the total number of TPRs proteome-wide. This study established that the minimum functional fold are 3-TPR-Domains.
3. Tetratricopeptide Repeat Family

3.2 Introduction

Figure 3.1: TPR sequence logo and structural features. (A) Sequence logo of the TPR MSA in Figure 3.5. Generated with WebLogo (Crooks et al., 2004) on the online webserver tool (https://weblogo.berkeley.edu/logo.cgi) The colour scheme was adopted from Jalview (Waterhouse et al., 2009) that emulates Clustal X (Larkin et al., 2007) (B) Example TPR from
CTPR1 (PDB code: 1NA0; Main et al., 2003), highlighting conserved sites from the literature. Positions 4, 7, 8, and 11 in Helix A (orange), positions 20, 24, 27 in Helix B (cyan), and position 32 (blue). (C) A TPR-Domain structure of OGT (PDB code: 1W3B; Jínek et al., 2004), where each TPR is coloured differently from adjacent TPRs. (D) A TPR-Domain structure example of PP5 (PDB code: 1A17; Das et al., 1998) to visualise the concave (blue) and convex (orange) surfaces of the TPR-Domain.

TPR domains have a convex and concave surface, as illustrated in Figure 3.1D. A variety of substrates, including RNA molecules, bind at the concave surface and, more rarely, the convex surface (Perez-Riba and Itzhaki, 2019, Zeytuni and Zarivach, 2012). Furthermore, most residues that interact with substrates on the concave surface are not conserved across different repeats (Magliery and Regan, 2005). This residue diversity reflects the observation that TPRs have evolved to accommodate for an extensive repertoire of substrates, ranging from peptides and disordered proteins to large globular domains (Perez-Riba and Itzhaki, 2019, Zeytuni and Zarivach, 2012).

The topological arrangement of TPR-Domains have often been compared to 14-3-3 domains (Xiao et al., 1995) and Sec7 domains (Mossessova et al., 1998, Cherfils et al., 1998). However, they are not assembled from canonical TPRs and they are relatively longer with varied helical packing (Groves and Barford, 1999).

TPRs belong to a superfamily that share sequence homology with four other evolutionarily related TPR-like repeat families that all share a helix-turn-helix structural motif. PPRs (Aubourg et al., 2000, Small and Peeters, 2000) are specialised in RNA-binding to regulate post-transcriptional gene expression in the nucleus and organelles (Small et al., 2013). Acknowledging the abundance of TPR-containing proteins in prokaryotes (Jernigan and Bordenstein, 2015), the high sequence similarity to TPRs meant it was believed PPRs may have emerged from TPRs during early eukaryote evolution (Reviewed in Barkan and Small, 2014). In addition, PPRs are have undergone considerable expansions within plants (Schmitz-Linneweber and Small, 2008). PPRs differs slightly at identical conserved sites in TPRs;
position 8 is conserved for aliphatic properties, instead of glycine/alanine in TPRs; position 17 and 24 are also conserved for aliphatic properties, instead of aromatic residues in TPRs; PPRs are additionally conserved for cysteine at position 12 and aspartate at position 34.

Unlike the PPR family, the SLR family of proteins spans a wide range of functions: Protein sel-1 homolog 1 (*Homo sapiens*) is involved with the degradation of misfolded endoplasmic reticulum, amongst other roles related to the pancreas (Lilley and Ploegh, 2005); during sporulation the Sporulation-specific homolog of CSD4 (SHC1) protein activates chitin synthase III, which synthesises chitin in cell walls of *Saccharomyces cerevisiae* (Sanz et al., 2002); Localization factor PodJL helps support the localisation of several polar organelles in *Caulobacter vibrioides* (Vioillier et al., 2002). SLRs may have developed in the last common ancestor between bacteria and eukaryotes, since they were absent in archaea (Mittl and Schneider-Brachert, 2007). While individual TPRs and SLRs reveals RMSD under 1Å upon superimposition, the SLR-Domain differs drastically from TPR-Domain. SLRs have a distinct L-shaped loop between Helix 1 and 2 that is 4–12 amino acids long but has a deletion of 2 amino acids in the second loop. The first loop introduces a large gap between adjacent SLRs, creating a distinctly different solenoid fold with slightly larger pitch by 1 or 2 SLRs more than TPRs.

TPRs can bind to a variety of substrates, from binding to individual proteins (Zhang et al., 2015) RNA strands (Katibah et al., 2014, Katibah et al., 2013) to organising large multi-protein complexes (Alfieri et al., 2017). There is a huge array of repeat structures solved with their substrates, highlighting the diversity of TPR proteins and their ability to bind to a wide range of substrates.

Perez-Riba and Itzhaki (2019) put forward five different bindings modes of the TPR family using structures archived in the PDBe (Velankar et al., 2010): short peptide binding to the concave surface; long extended peptide binding to the concave surface; TPR hinge
formation; binding of folded domains to TPR domains; homo-dimers, homo-oligomers, and domain-swap. Short peptide binding is dominated by 3-TPR-Domains from proteins like Hsp70 and Hsp90 organizing protein (HOP) (1ELR), Small glutamine-rich tetratricopeptide repeat-containing protein 2 (Sgt2) (5LYN), and PP5 (2BUG) are some of many proteins that bind to the MEEVD motif of Hsp70 and 90 proteins (Scheufler et al., 2000). Long extended concave peptide binding includes G-protein-signalling modulator 2 (GPM2) (3SF4) and CTR9 that binds to two proteins simultaneously (6AF0). PTS1R and its homologues are the only ones that fold on themselves with a TPR hinge formation to bind to its target. Lastly, Lipopolysaccharide assembly B (LapB) form homo-dimers, while Sorting nexin-21 (SNX21) (4YMR) and IFIT2 (4G1T) display domain-swapping. The final and undiscussed arrangement and includes both long convex peptide binding as well as simultaneous concave and convex binding with extended proteins. The TPR-containing proteins making up the APC primarily employ convex binding (4UI9). Multi-protein complexes like Superkiller protein 3 (SKI3) have long peptides that have extended concave binding and convex binding (4BUJ).

3.3 Methods

3.3.1 Gathering repeat annotations

Sampling TPR-containing proteins in InterPro, the InterPro Entry that best annotates on the Repeat Biological Entity is IPR019734. The contributing member databases of this InterPro Entry are SMART (SM00028), PROSITE (PS50005), and Pfam (PF13176, PF13181, PF13174). Furthermore, all TPR annotations from Swiss-Prot with the keyword "TPR repeat [KW-0802]" were also retrieved. Figure 3.2 shows the comparison of how the different databases annotate TPRs in proteins from Homo sapiens. A non-redundant merge of all in Figure 3.3A these annotations from the databases in the order SMART, PROSITE, Pfam, and Swiss-Prot, produced a total of 1028 TPRs, termed Non-Redundant TPRs. Non-Redundant TPRs comprise 649 SMART, 33 PROSITE, and 346 Swiss-Prot annotations. There were no
Pfam annotations as SMART and PROSITE annotated the same regions as them. This dataset spans 166 human TPR-containing proteins.

Figure 3.2: Comparison of TPR annotations from four databases. UpSet plots visualising how Pfam (PF13174, PF13176, PF13181), PROSITE (PS50005), SMART (SM00028), and Swiss-Prot (“TPR”) databases annotate TPRs.

The PRA of Non-Redundant TPRs is fragmented since there are gaps between TPRs comparable to that of canonical TPRs. Furthermore, sampling TPR-containing proteins showed canonical TPRs visible in three-dimensional structures of some proteins that were not captured by this method. In the interest of studying full-length TPR-Domains, the protocol described in Figure 3.3 was followed to identify TPRs currently not annotated in Swiss-Prot or InterPro.
3. Tetratricopeptide Repeat Family

3.3 Methods

Figure 3.3: Reannotating human TPRs with maxHMM. The four steps include: (A) Gathering and perform non-redundant merge of TPRs, (B) Multiple sequence alignment of Seed TPRs and filtering atypical lengths, (C) Building a Multi-level HMM library and the assessment of its performance, (D) Automatic and manual curation of the Final TPRs and aligning them.

3.3.2 Multiple sequence alignment

In Figure 3.3B, Non-Redundant TPRs were filtered by repeat length to remove TPRs < 31 or > 35 amino acids. This step left 927 Seed TPRs and 101 Filtered TPRs, which are revisited.
3.3 Methods

later in Figure 3.3D. The purpose of this step was to remove TPRs with large indels that sometimes deviated from the consensus sequence and thus, complicated the alignment stage. The major difficulty with aligning non-canonical TPRs is that alignment methods would introduce insertions within Helix A and B.

Seven different alignment methods were assessed with the TPR MSA; ClustalW, MAFFT, T-Coffee, ProbCons, Clustal, GLProbs (Ye et al., 2014), and MUSCLE. Table 3.1 lists the MSA properties built with all the alignment methods. Since the MSAProbs method (Version 0.9.7; Liu et al., 2010) performed the best, this MSA and all other global alignments of TPR sequences later in this chapter have used this method. Alignment method parameters included 500 passes of iterative refinement and five passes of consistency transformation.

Table 3.1: The performance of eight different global alignment methods assessed on Seed TPR sequences. Each alignment program used here had default parameters. Alignment programs have been sorted in descending order of performance by the number of gap-free sequences they generate. The rest of the sequences not numbered in the table are the 17 long and non-canonical TPR sequences that are always occupying low-occupancy columns.

<table>
<thead>
<tr>
<th>Sequence Alignment Program</th>
<th>Gap-Free Sequences</th>
<th>Gap-Occupying Sequences</th>
<th>Low Occupancy Columns</th>
<th>Short Length Sequences</th>
<th>Canonical Length Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSAProbs</td>
<td>632</td>
<td>17</td>
<td>35</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ProbCons</td>
<td>631</td>
<td>18</td>
<td>24</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Clustal Omega</td>
<td>630</td>
<td>19</td>
<td>22</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>GLProbs</td>
<td>626</td>
<td>23</td>
<td>43</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>625</td>
<td>24</td>
<td>31</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Clustal W</td>
<td>351</td>
<td>298</td>
<td>41</td>
<td>6</td>
<td>275</td>
</tr>
<tr>
<td>MAFFT</td>
<td>344</td>
<td>305</td>
<td>149</td>
<td>8</td>
<td>280</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>309</td>
<td>340</td>
<td>41</td>
<td>3</td>
<td>320</td>
</tr>
</tbody>
</table>

Sequence Alignment Program: The name of the alignment program. Gap-Free Sequence(s): The number of sequences that solely contribute to alignment columns with 50% or more residue occupancy in the alignment. Gap-Occupying Sequence(s): The number of sequences that have at least one residue contribute to alignment columns with less than 50% residue occupancy in the alignment. Low-Occupancy Columns: The number of alignment columns that have less than 50% residue occupancy in the alignment. Short Length Sequences: The number of gap-occupying sequences that are shorter than the canonical length of TPRs. Canonical Length Sequences: The number of gap-occupying sequences that have the canonical length of TPRs.
The alignments were assessed by how well they aligned to conserved positions in TPRs (see Figure 3.1), introducing the fewest gaps in the helices. Accordingly, the MSAProbs aligned MSA with 927 Seed TPRs was then manually curated to correct for the 17 long and non-canonical TPRs that occupy low-occupancy columns in Table 3.1.

3.3.3 Multi-level HMM library

Firstly, the performance of a single HMM profile was assessed as a reference point to justify the use of maxHMM. The single HMM profile will then be compared with the multi-level HMM library of maxHMM. All HMM profiles were then searched against full-length human protein sequences in the Swiss-Prot database. Resulting hits were classified as True Positive if they matched a Seed TPR, True Negative if they matched a region that are not TPR annotations in InterPro (See Section 2.3.6.3), or Other matches if they matched neither.

3.3.3.1 Single HMM-profile

An HMM profile was built from the Seed TPRs MSA. With the default sequence bit score inclusion threshold of 10, Figure 3.4 shows the HMM profile had poor discrimination between Seed TPRs and known Non-TPRs.
3.3 Methods

Figure 3.4: Performance of a single HMM profile against maxHMM (multi-level HMM library). Both histograms visualise three types of hits: Seed TPRs, in green; True Negative, in red; All, matches that are neither Seed TPRs nor True Negative. True Positive matches are the Seed TPR annotations with which the HMM-Profiles were trained on. True Negative matches are the annotations that are Non-TPRs. (A) Single HMM-Profile and (B) maxHMM.

3.3.3.2 MaxHMM

Figure 3.3 shows the stages in building the multi-level HMM library, described in detail in Section 2.3.6. AMPS calculate Z-scores from pairwise sequence alignments from the TPR MSA. OC program built a tree with complete linkage clustering using Z-scores and the Seed TPR alignment was dissected at increasing Z-scores to give 8, 31, and 69 clusters. HMM profiles were built from the full alignment and all clusters to create a library of 109 HMM profiles that were searched against reviewed full-length proteins in *Homo sapiens*, resulting in a total of 40319 raw hits.

The resulting 2378 matches were labelled if they overlapped with the Seed TPRs trained on or the 103828 true negative InterPro annotations. Plotting the bit scores for all hits,
Figure 3.4B illustrates the improved discrimination obtained by the multi-level HMM library. At an optimal bit score threshold of 19, 925/927 Seed TPRs recalled with a false positive rate of 5/103828 scores a Matthews correlation coefficient (MCC) of 0.997, compared to an MCC score of 0.901 for a single HMM at a bit score threshold of -4.

The workflow in Figure 3.3D splits as the filtered dataset of hits was split by the threshold into 969 hits scoring above the threshold, and 1409 below. The next stages of Figure 3.3D show each subdivided set of hits dissects further that recognise Seed TPRs, Non-TPRs, and Other hits recognising neither.

Of the 925 Seed TPRs that met the threshold, 113 Seed TPRs were discarded for multiple reasons. Many of them were Swiss-Prot annotations, incorrectly labelled as a TPR, that annotated TPR-like regions. In other cases, some proteins had sparsely annotated TPRs that were spaced too far apart to recapitulate a complete TPR-Domain. In the interest of studying intact TPR-Domains that comprise repeat annotations with clear sequence homology to canonical TPRs, all TPR annotations from TPR-containing proteins with incomprehensive annotations were discarded entirely.

Scoring below the threshold, there are many Other matches that often resided at regions with long series of helix-turn-helix, annotated by InterPro entries ‘Tetratricopeptide repeat-containing domain’ (IPR013026) and the ‘Tetratricopeptide-like helical domain superfamily’ (IPR011990). IPR013026 and IPR011990 are capable of annotating TPR-Domains and TPR-like regions with high confidence, respectively.
Table 3.2 counts all the different matches and their labels. Values above a count of a hundred will be focused on here to demonstrate the trends observed. The 201 Other matches that are isolate are very likely helix-turn-helix motifs in a domain. There are 129 Other matches that are adjacent to Other matches, which are likely TPR. Although not shown in
Table 3.2, there are also 129 Non-TPRs matches that are in tandem. A sizeable proportion of these matches are in fact TPR-like regions. This is because the True Negative dataset also includes annotations that belong to the TPR Clan (CL0020) in Pfam, to help dissect canonical TPRs from TPR-like-containing proteins discussed earlier in Section 3.2. Pfam Clans include Pfam families that share sequence similarities that may have originated from a common ancestor, have structural similarity to known structures, and functional similarity (Finn et al., 2006). Lastly, Other matches that are of most interest are the 134 matches that overlap with TPR-like superfamily region (IPR011990) and is in tandem with Other matches. These matches tend to represent regions that are canonical TPRs.
Table 3.2: The number of Other matches scoring below the bit score threshold and their labels. There are four labels: matches overlap with Tetratricopeptide-like helical domain superfamily’ (IPR011990; CATH-Gene3D database); Repeat Region includes Seed TPRs and ‘Tetratricopeptide repeat-containing domain’ (IPR013026); PRA analysis identifies whether Other matches are isolate or in tandem with another annotation(s), such as Other, TPR, or Non-TPR.

<table>
<thead>
<tr>
<th>CATH</th>
<th>Repeat Region</th>
<th>Tandem PRA</th>
<th>Tandem PRA Categories</th>
<th>Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>Isolate</td>
<td>-</td>
<td>201</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>Tandem</td>
<td>All Clashing</td>
<td>2</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>Tandem</td>
<td>TPR and Other</td>
<td>39</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>Tandem</td>
<td>Other</td>
<td>129</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>Tandem</td>
<td>Non-TPR and Other</td>
<td>11</td>
</tr>
<tr>
<td>-</td>
<td>TPR</td>
<td>Isolate</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>TPR</td>
<td>Tandem</td>
<td>TPR and Other</td>
<td>15</td>
</tr>
<tr>
<td>-</td>
<td>TPR</td>
<td>Tandem</td>
<td>Other</td>
<td>6</td>
</tr>
<tr>
<td>CATH</td>
<td>-</td>
<td>Isolate</td>
<td>-</td>
<td>46</td>
</tr>
<tr>
<td>CATH</td>
<td>-</td>
<td>Tandem</td>
<td>All Clashing</td>
<td>5</td>
</tr>
<tr>
<td>CATH</td>
<td>-</td>
<td>Tandem</td>
<td>TPR and Other</td>
<td>134</td>
</tr>
<tr>
<td>CATH</td>
<td>-</td>
<td>Tandem</td>
<td>Other</td>
<td>116</td>
</tr>
<tr>
<td>CATH</td>
<td>-</td>
<td>Tandem</td>
<td>Non-TPR and Other</td>
<td>6</td>
</tr>
<tr>
<td>CATH</td>
<td>TPR</td>
<td>Isolate</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>CATH</td>
<td>TPR</td>
<td>Tandem</td>
<td>All Clashing</td>
<td>3</td>
</tr>
<tr>
<td>CATH</td>
<td>TPR</td>
<td>Tandem</td>
<td>TPR and Other</td>
<td>95</td>
</tr>
<tr>
<td>CATH</td>
<td>TPR</td>
<td>Tandem</td>
<td>Other</td>
<td>13</td>
</tr>
</tbody>
</table>

While many Other matches recognised these regions, the focus of this step in the workflow of Figure 3.3D was to rectify TPR-Domains that are missing TPR annotations. Manual curation resolved 183 Other hits as genuine TPRs.

Following the manual curation of all hits in target proteins, a total of 1140 hits were discarded in the process. Earlier in the workflow of Figure 3B, 101 Filtered TPRs were removed from Non-Redundant TPRs before aligning. Only 15 Filtered TPRs remained following manual curation, which were not detected by the Multi-level HMM method.
This workflow successfully filled the gaps in TPR-Domains, where appropriate, which resulted in 193 previously unannotated TPRs. The manual curation step in Figure 3D identified 33 new TPRs discovered through sequence, repeat architecture, and structure analysis that the multi-level HMM profile library did not detect. Thus, contributing to the complete and Final TPR dataset of 1072 TPRs across 145 TPR-containing proteins. These TPRs were then aligned in the same manner as the Seed TPRs, earlier in Figure 3B.

3.3.3.3 Comparison to other databases

To confirm that the Final TPRs are a comprehensive set of annotations, 276 TPRs observed in structures match with 61/65 TPRs from RepeatsDB, as defined by the Class-Topology-Fold-Clan classification 3.3.2.1 (Paladin et al., 2021). The 4 TPRs absent in the Final TPRs represent single helices or truncated TPRs that are not full-length N- or C-terminus capping TPRs, but solvating helices shielding the hydrophobic regions of a TPR-Domain. Thus, they were ignored and the Final TPR was considered comprehensive to be used throughout this study.

3.4 Results and discussion

3.4.1 Repeat architecture

The organisation of TPRs was last analysed by D'Andrea and Regan in 2003. Since then, it has set the core background literature known today. While this review has established the arrangement of TPRs, it has three limitations today; D'Andrea and Regan used TPRs from the SMART database with redundant set of annotations, due to overcounting TPRs from protein isoforms; the analysis is very primitive since TPRs can accommodate much more diversity in repeat organisation, highlighted later in this chapter; this analysis is from 18 years ago.

The SMART dataset includes isoforms and lists identical TPR annotations across isoforms. While not explicitly stated in the review by D'Andrea and Regan (2003), nor communicated through personal communication, it can be safely assumed the dataset they used
included TPR annotations from isoforms. This assumption is based on how similar the number of annotations used in their study (1492 TPRs in 260 proteins) is to the number of the confirmed redundant annotations in the SMART database today (1729 annotations in 300 proteins). Furthermore, their analysis refers to the total number of TPRs in each protein. Although this is a useful property to measure, which is included in the repeat architecture analysis of RAM, it is an oversimplification of repeat organisation.

The RepeatsDB also studied the PRA of the TPR family through their summary statistics page. It is not biologically relevant as it captures the total number of TPR-Domains across a redundant collection of structures across all TPR-containing proteins. Instead, it is more useful to capture what structures the PDBe archives for this classification. Furthermore, a key limitation of RepeatsDB today, comparable to Swiss-Prot in its infancy, is that this database is manually curated and has a backlog of protein entries yet to be reviewed.

Figure 2.5 visualises the updated statistics of TPRs in humans with a more sophisticated repeat architecture analysis. Figure 2.5A shows the most frequent number of TPRs is 3, regardless of their position in the protein. That figure was an assurance as it agrees with observations in D’Andrea and Regan (2003). Assuming a spacing threshold of half the length of a canonical TPR at 17 amino acids, Figure 2.5B shows the most common tandem TPR arrangements form 2-TPR-Domains or 3-TPR-Domains. This part of the analysis was a clearer way to confirm the current dogma that 3-TPR-Domains are the shortest functional TPR-Domains. Collectively this organisation contributes to a total of 124/220 (56.6%) 2-TPR-Domains and 3-TPR-Domains observed in all TPR-containing proteins. As proteins can have multiple TPR-Domains, Figure 2.5C shows that proteins can have as many as five TPR-Domains, while the most common is one.

The repeat architecture analysis so far deconstructs the organisation of human TPRs proteome-wide. To combine all the information together and understand how TPR-containing
proteins employ and organise multiple TPR-Domains, Figure 2.5D shows how frequent each combination of TPR-Domains is observed in TPR-containing proteins. Nearly half of all 3-TPR-Domains (39/83; 47.0%) are the only ones present in TPR-containing proteins. Whereas other proteins, with two or more TPR-Domains, tend to have 3-TPR-Domains paired with other TPR-Domains with no apparent logic. Although this has not been discussed in the TPR literature, it is surprising not to observe longer TPR-Domains organised in multiples of three, like in the Ankyrin and LRRs families (Paladin et al., 2020). Comparable to HEATs, many TPR-containing proteins may have had a longer evolutionary history and expanded before speciation. Thus, masking the original exon junctions that have brought about the current organisation of TPRs (Schaper et al., 2014).

3.4.2 Sequence analysis
The next section entails analysing the whole alignment by using standard methods for scoring residue conservation and this will be supplemented by summarised physicochemical properties. This analysis extends towards subfamily analysis that splits the alignment on two accords: general repeat architecture across all TPR-Domains that creates start, middle, and end TPR MSAs; functional TPRs that make up 3-TPR-Domains.

3.4.2.1 Full alignment
Visualised onto the structure of Figure 3.6 and highlighted above the MSA (green) in Figure 3.5, the literature has previously established positions 4, 7, 8, 11, 20, 24, 27, and 32 as the conserved residues making up TPRs. Denoted above the MSA in Figure 3.5, 6/8 positions (red) from the literature are below PRShenkin of 25%. Those positions are visualised onto the structure of Figure 3.6. Summarised in Table 3.3, these sites are conserved for residue identity, where position 8 and 20 are conserved for alanine/glycine and glycine, respectively.
3. Tetratricopeptide Repeat Family

3.4 Results and discussion

Figure 3.5: TPR MSA. The 27 sequences shown represent equidistant members of an average distance tree, calculated in Jalview from the full alignment. All annotations above and below the MSAs refer to information from the full-length MSA with 1028. Annotation above MSA: The annotation ladder (black) numbers the alignment columns in intervals of 5 amino acids. Low occupancy alignment columns, less than 50% residue occupancy, are hidden. Arrays of hidden columns are collapsed and depicted as a single blue line through the alignment with an arrowhead above the alignment. Annotation below MSA: two annotation tracks below the alignment that summarises the information for the full-length MSA; PR_{Shenkin}, where lower scores indicate more conserved positions, and vice versa for larger values; secondary structure allocated based on summarised DSSP definitions across all human TPR structures in Figure 3.9.

Table 3.3: Top three most frequent amino acids in each alignment column of the TPR MSA in Figure 3.5. The percentage values represent the proportion of the amino acid present in an alignment column.
### Alignment Column: High-occupancy alignment column in the MSA of Figure 3.5. Rank: The top three most frequency amino acid at each alignment position.

<table>
<thead>
<tr>
<th>Rank</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alanine (23.9%)</td>
<td>Valine (9.5%)</td>
<td>Leucine (8.8%)</td>
</tr>
<tr>
<td>2</td>
<td>Glutamate (13.4%)</td>
<td>Lysine (11.7%)</td>
<td>Arginine (8.7%)</td>
</tr>
<tr>
<td>3</td>
<td>Alanine (24.6%)</td>
<td>Leucine (15.3%)</td>
<td>Valine (11.1%)</td>
</tr>
<tr>
<td>4</td>
<td>Leucine (27.2%)</td>
<td>Tyrosine (16.8%)</td>
<td>Tryptophan (7.9%)</td>
</tr>
<tr>
<td>5</td>
<td>Leucine (16.4%)</td>
<td>Tyrosine (10.9%)</td>
<td>Phenylalanine (8.5%)</td>
</tr>
<tr>
<td>6</td>
<td>Asparagine (21.1%)</td>
<td>Leucine (11.2%)</td>
<td>Arginine (8.9%)</td>
</tr>
<tr>
<td>7</td>
<td>Leucine (39.8%)</td>
<td>Arginine (11.8%)</td>
<td>Isoleucine (7.1%)</td>
</tr>
<tr>
<td>8</td>
<td>Alanine (41.2%)</td>
<td>Glycine (41.1%)</td>
<td>Serine (4.9%)</td>
</tr>
<tr>
<td>9</td>
<td>Leucine (13.0%)</td>
<td>Asparagine (9.7%)</td>
<td>Lysine (8.5%)</td>
</tr>
<tr>
<td>10</td>
<td>Leucine (19.2%)</td>
<td>Alanine (16.3%)</td>
<td>Valine (10.4%)</td>
</tr>
<tr>
<td>11</td>
<td>Tyrosine (32.3%)</td>
<td>Leucine (22.5%)</td>
<td>Histidine (8.0%)</td>
</tr>
<tr>
<td>12</td>
<td>Leucine (17.8%)</td>
<td>Tyrosine (10.8%)</td>
<td>Phenylalanine (9.9%)</td>
</tr>
<tr>
<td>13</td>
<td>Lysine (15.6%)</td>
<td>Alanine (10.6%)</td>
<td>Arginine (10.6%)</td>
</tr>
<tr>
<td>14</td>
<td>Leucine (24.7%)</td>
<td>Glutamine (11.3%)</td>
<td>Glutamate (7.7%)</td>
</tr>
<tr>
<td>15</td>
<td>Glycine (43.3%)</td>
<td>Lysine (11.9%)</td>
<td>Glutamate (8.7%)</td>
</tr>
<tr>
<td>16</td>
<td>Aspartate (15.7%)</td>
<td>Lysine (14.3%)</td>
<td>Asparagine (13.7%)</td>
</tr>
<tr>
<td>17</td>
<td>Tyrosine (26.8%)</td>
<td>Leucine (13.6%)</td>
<td>Phenylalanine (12.7%)</td>
</tr>
<tr>
<td>18</td>
<td>Glutamate (22.3%)</td>
<td>Aspartate (12.4%)</td>
<td>Lysine (11.0%)</td>
</tr>
<tr>
<td>19</td>
<td>Glutamate (26.3%)</td>
<td>Lysine (14.6%)</td>
<td>Glutamine (8.5%)</td>
</tr>
<tr>
<td>20</td>
<td>Alanine (79.5%)</td>
<td>Serine (5.0%)</td>
<td>Valine (3.9%)</td>
</tr>
<tr>
<td>21</td>
<td>Leucine (25.9%)</td>
<td>Isoleucine (15.6%)</td>
<td>Valine (11.4%)</td>
</tr>
<tr>
<td>22</td>
<td>Glutamate (18.3%)</td>
<td>Lysine (13.4%)</td>
<td>Glutamine (9.9%)</td>
</tr>
<tr>
<td>23</td>
<td>Leucine (13.8%)</td>
<td>Tyrosine (11.8%)</td>
<td>Cysteine (10.3%)</td>
</tr>
<tr>
<td>24</td>
<td>Tyrosine (26.3%)</td>
<td>Leucine (23.8%)</td>
<td>Phenylalanine (19.1%)</td>
</tr>
<tr>
<td>25</td>
<td>Glutamate (18.8%)</td>
<td>Lysine (13.0%)</td>
<td>Glutamine (12.1%)</td>
</tr>
<tr>
<td>26</td>
<td>Lysine (20.4%)</td>
<td>Arginine (16.8%)</td>
<td>Glutamate (15.4%)</td>
</tr>
<tr>
<td>27</td>
<td>Alanine (49.5%)</td>
<td>Leucine (9.3%)</td>
<td>Valine (8.9%)</td>
</tr>
<tr>
<td>28</td>
<td>Leucine (43.6%)</td>
<td>Isoleucine (13.5%)</td>
<td>Valine (9.0%)</td>
</tr>
<tr>
<td>29</td>
<td>Glutamate (18.5%)</td>
<td>Lysine (14.8%)</td>
<td>Glutamine (11.4%)</td>
</tr>
<tr>
<td>30</td>
<td>Leucine (31.0%)</td>
<td>Isoleucine (12.6%)</td>
<td>Valine (6.9%)</td>
</tr>
<tr>
<td>31</td>
<td>Aspartate (13.4%)</td>
<td>Asparagine (11.7%)</td>
<td>Alanine (11.3%)</td>
</tr>
<tr>
<td>32</td>
<td>Proline (35.2%)</td>
<td>Glutamate (8.2%)</td>
<td>Lysine (7.1%)</td>
</tr>
<tr>
<td>33</td>
<td>Aspartate (11.7%)</td>
<td>Glutamate (9.9%)</td>
<td>Lysine (9.8%)</td>
</tr>
<tr>
<td>34</td>
<td>Asparagine (14.6%)</td>
<td>Leucine (10.8%)</td>
<td>Aspartate (10.0%)</td>
</tr>
</tbody>
</table>
Other conserved positions outside of the literature set in the MSA (Figure 3.5) include positions 15 and 28. Figure 3.6A shows position 15, located at the first turn following the end of Helix A. That position is dominated by glycine (Table 3.3) with 43.3% occupancy. This site likely promotes flexibility in the turn to help the two helices fold on each other. Like position 32 with proline (35.2% occupancy), this site may act as a helix breaker. Together, these two sites define the boundaries of a TPR and the helices within. Position 28 is another conserved position, in Figure 3.6A, that packs against Helix A of the next TPR in the N-terminus direction. This site comprises leucine (43.6%), isoleucine (13.5%), and valine (9.0%) amino acids.

TPRs have very few positions (8/34) that are conserved, using standard methods for scoring residue conservation. TPRs have often been qualitatively described as having additionally conserved sites based on the physicochemical properties of amino acids at each position. For example, position 24 has large and aromatic sidechains (Figure 3.1; Zeytuni and Zarivach, 2012). To quantitatively measure this property,
Table 3.4 summarises the top three physicochemical properties at each position.
Table 3.4: Top three most common physicochemical properties at each alignment position for the TPR MSA in Figure 3.5. The physicochemical properties uses the amino acid property index, as described by Livingstone and Barton (1993) from the Venn diagram defined by Taylor (1986). The physicochemical properties were obtained for each amino acid in the MSA. The properties were then summed for each alignment column and sorted in descending order of frequency, to show the top three most common properties.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Alignment Column</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Tiny</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Charged</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Tiny</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Hydrophobic</td>
<td>Polar</td>
<td>Aromatic</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Hydrophobic</td>
<td>Polar</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Polar</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Small</td>
<td>Hydrophobic</td>
<td>Tiny</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Aliphatic</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Hydrophobic</td>
<td>Polar</td>
<td>Aromatic</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Hydrophobic</td>
<td>Polar</td>
<td>Aliphatic</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Polar</td>
<td>Hydrophobic</td>
<td>Charged</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Hydrophobic</td>
<td>Polar</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Polar</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Polar</td>
<td>Charged</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Hydrophobic</td>
<td>Polar</td>
<td>Aromatic</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Polar</td>
<td>Charged</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Polar</td>
<td>Charged</td>
<td>Hydrophobic</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Small</td>
<td>Hydrophobic</td>
<td>Tiny</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Polar</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Polar</td>
<td>Charged</td>
<td>Hydrophobic</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Hydrophobic</td>
<td>Polar</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Hydrophobic</td>
<td>Aromatic</td>
<td>Polar</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Polar</td>
<td>Charged</td>
<td>Hydrophobic</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Polar</td>
<td>Charged</td>
<td>Hydrophobic</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Hydrophobic</td>
<td>Small</td>
<td>Tiny</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Polar</td>
<td>Charged</td>
<td>Hydrophobic</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Hydrophobic</td>
<td>Aliphatic</td>
<td>Polar</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Polar</td>
<td>Small</td>
<td>Hydrophobic</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Small</td>
<td>Polar</td>
<td>Proline</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Polar</td>
<td>Small</td>
<td>Hydrophobic</td>
<td></td>
</tr>
</tbody>
</table>
3. Tetratricopeptide Repeat Family

3.4 Results and discussion

<table>
<thead>
<tr>
<th></th>
<th>Polar</th>
<th>Small</th>
<th>Hydrophobic</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alignment Column: High-occupancy alignment column in the MSA of Figure 3.5. Rank: The top three most common physicochemical properties at each alignment position.

Studying the physicochemical properties in parallel with the most conserved positions, there are positions that share similar combinations. All seventeen positions were split into four classes with hydrophobic sidechains: small and tiny, in Figure 3.6B; small and aliphatic, in Figure 3.6C; aliphatic and polar, in Figure 3.6C; aromatic and polar, in Figure 3.6D; polar and small, in Figure 3.6E.

Evolutionary conserved positions 8, 20 and 27 are paired with positions 1 and 3. In Table 3.3, these two sites share the same top three amino acids with position 27. Unlike positions 8 and 20, positions 1 and 3 comprise alanine, with 23.9% and 24.6%, valine (9.5/15.3%), and leucine (8.8/11.1%). Visualising them onto the CTPR protein in Figure 3.6B, together positions 1, 3, and 27 look like they have more clearance for larger sidechains.
3. Tetratricopeptide Repeat Family

3.4 Results and discussion

Figure 3.6: Evolutionary conserved positions in TPRs. Two adjacent TPRs of the consensus designed CTPR protein (PDB code: 1NA0; Main et al., 2003) coloured by evolutionary conserved proteins and other positions which are conserved for physicochemical properties. (A) The evolutionary conserved positions with the lowest PRShenkin (<=25%) in the TPR MSA of Figure 3.5, were coloured and their sidechains displayed. Residues that remain conserved and consistent with the literature were coloured in red and its sidechains displayed. This threshold did not recapitulate positions 4 and 32 from the literature and was not conserved (grey). Positions 15 and 28 are conserved outside of the literature (green). Conserved physicochemical properties of an alignment position, obtained from the amino acid property index defined in Livingstone and Barton (1993). Positions that were hydrophobic were further classified into five classes: (B) small and tiny, in red; (C) small and aliphatic, in orange; (D) aliphatic and polar, in teal; (E) aromatic and polar, in blue; (F) polar and small, in green.

The next set includes slightly larger sidechains. Coupled with the conserved position 28, Table 3.3 shows position 10 has a much more even split between alanine (19.2%), valine (16.3%), and leucine (10.4%). Position 28 contrasts with a significantly larger proportion of amino acids with aliphatic properties. Studying these two sites in Figure 3.6C, this set begins to point away from the neighbouring helices, with more room to accommodate for slightly larger sidechains.

Like the previous set, this next set of positions with aliphatic and polar sidechains (Figure 3.6D) continue to steer away from the helices and the hydrophobic pockets between them. Thus, longer, and larger sidechains form aliphatic sidechains, while being more solvent, with polar properties. Positions 21 and 30 are near identical to position 28 in the amino acid composition. The conserved position 7 is a slight exception to this set that has arginine (11.8%), in Table 3.3, instead of valine in the top three amino acids. The biggest outlier in this set is position 12, which sits between this and the next set. Although this site has leucine (17.8%), the next two top amino acids are large and aromatic amino acids tyrosine (10.8%) and phenylalanine (9.9%).

The next set characteristically have the largest set of sidechains that are aromatic and polar. Paired with conserved sites 11 and 24, positions 4 and 17 in Figure 3.6E show they are in regions that represent the clefts between the same helices of different TPRs. For example,
positions 4 and 11 makes up the hydrophobic pocket between Helix B and B’ of TPR and TPR’, respectively.

Lastly, the final set contrast from the other sets in that they are small and polar. This is likely due to the presence of polar amino acids like asparagine and glutamine across most sites. The two exceptions are positions 15 and 32. In Table 3.3, position 15 is evenly split between glycine (43.3%) and other charged residues, creating split properties between small and polar properties.

Position 32 is a standalone site, as it has the highest proportion of proline (35.2%). All the other sites not discussed already are positions with charged properties. These will be covered later when discussing how TPRs are involved with substrate interaction.

3.4.2.2 Subfamily analysis of 3-TPR-Domains

Discussed earlier in Section 1.3.5.3.2, clustering sequences and cutting the resulting dendrogram highlights three large subfamilies. Each subfamily represents individual repeats that make up a 3-TPR-Domain: Figure 2.12A, 3-TPR-1-Domain, Figure 2.12B, 3-TPR-2-Domain, Figure 2.12C, 3-TPR-3-Domain. These 3-TPR-Domains are conserved and specialised for binding to the MEEVD motif of HSP70/90 proteins. The additional conserved sites outside of the ones discussed already include position 5, 6, 9, and 13, which are involved with binding to the motif (Scheufler et al., 2000) and commonly employed by TPR-Domains (Magliery and Regan, 2005).

3.4.2.3 Splitting the alignment by repeat architecture

Initially introduced in Section 1.3.5.3.1 in Methods, TPRs can be classified further based on their general position in the TPR-Domains. TPRs at the start and end of TPR-Domains are likely to have different conservation profiles to TPRs in the middle since some positions are solvent-accessible. There are 220 start and end TPRs, 561 middle TPRs. There are also 79 TPRs that do not belong to any TPR-Domain, but some have been observed in structure to fold with a TPR-Domain. Due to the lack of structural data for some of those TPRs, they were
ignored in the analysis. The alignment was then split into three separate alignments, as highlighted Figure 2.11, to study how the conservation profiles differ between start, middle, and end TPRs.

Building on further from the sequence analysis earlier, there are two pairs of sites that differ significantly in sequence conservation between different repeat architectures: 3 and 10, 12 and 17. Reflecting on the summarised physicochemical properties across the whole alignment in Figure 3.5, positions 3, 10, and 17 belong in the classes they have been assigned. However, position 12 becomes aromatic and polar in start TPRs (not shown). In contrast, this site also characterises as aliphatic and polar, and polar and small in middle and end TPRs, respectively. These two pairs of sites are unique to the repeat architecture and delineate further positions within a TPR-Domain.

Positions 7, 8, 11, 15, 20, 24, 27, and 28 in this study are recognised as the set of residues important for maintaining the characteristic fold in the TPR family. These conserved set of positions were further explored if they were also constrained in the human population. Additionally, positions 3, 10, 12, and 17 that differ in conservation between repeat architecture types were also investigated.

3.4.3 Variant analysis
Since the first and last TPR in a TPR-Domain have different solvent environments, they were considered separately in the analysis. Figure 3.7 shows the relationship between the MES and PR\textsubscript{Shenkin} for each column across all three TPR alignments. The analysis is split into two types: static positions, variable positions. Static positions do not change much in MES and PR\textsubscript{Shenkin} between different architectures. Variable positions have significant changes that can sometimes change their classifications altogether. Highlighted in Table 3.5 are differences in the two properties between start and end TPRs.
Across all three TPR architectures, position 24 is classified as a CMD in start ($MES = -0.35, p = 0.02$), middle ($MES = -0.24, p = 0.006$), and end ($MES = -0.18, p = 0.2$) TPRs. Unlike the most conserved positions 8 and 20, CMD site 24 remains the most depleted.

Contrasting CMD site 24, conserved positions 8, 7, 20, 27, 28, and 11 alternate between missense depleted (CMD) and enriched (CME) classifications. They lie within the margins of the confidence intervals (grey outline) of the modelled regression line.

From the top three largest changes scaled MES (Table 3.5), is the conserved position 10. Positions 3 and 10 were paired earlier in the sequence analysis for having one of the largest changes in $PR_{Shenkin}$ of 58.8% and 64.7%, respectively. This considerable difference is also reflected in the MES score with a decrease in MES of 0.6 (position 3) and 0.8 (position 10), from start to end TPRs. As a result, CME sites 3 and 10, in start TPRs, reclassify as CMD sites in end TPRs, where position 3 is marginally conserved. However, the large change in sequence conservation between start and middle TPRs, these sites are not missense depleted in middle TPRs.

Contrasting that pair are positions 12 and 17. Position 17 is classified as a CMD site in start TPRs, whereas position 12 only missense depleted. Being the next top two positions displaying significant changes in sequence conservation, they resemble the previous set of positions. Position 12 transitions ($MES = -0.15$) from missense depleted in start TPRs to CME in end TPRs. Whereas position 17 declassifies from a CMD site in start TPRs to neither missense enriched nor depleted in end TPRs.
3. Tetratricopeptide Repeat Family

3.4 Results and discussion

A

Start

B

Middle

C

End
Figure 3.7: Missense enrichment plots for TPR MSAs split by repeat architecture. Relationship between MES and sequence conservation. Regression model $y \sim x$ was fitted and the resulting regression line (black) with a 95% CI (grey outline) for that regression. The numbers at each datapoint refers to the column positions in the TPR alignments of Figure 2.11. Alignment column datapoints that meet the residuals threshold of 0 and reside in the lower (25%) and upper (75%) quartile $PR_{Shenkin}$ have been classified into four categories: Conserved and Missense Depleted (CMD) in blue, Conserved and Missense Enriched (CME) in red, Unconserved and Missense Depleted (UMD) in teal, Unconserved and Missense Enriched (UME) in yellow. Error bars indicate 95% CI.

The next largest change in Table 3.5 is position 15. In start TPRs, position 15 classifies as a CMD site and then transitions to a CME site in end TPRs. With very little change in conservation across all three repeat architectures. The variant analysis indicates this site is more important in start TPRs than in end.

Lastly, position 34 is the third largest change in MES. Since UMD sites are enriched for substrate interactions (MacGowan et al., 2017), this was investigated across all repeat architectures. Position 34 is marginally classified as UME in start TPRs. But it reclassifies as a UMD in both middle and end TPRs. Other UMD sites, in middle TPRs includes 5, 9, 12, and 13. Together, they all make up the concave surface involved with substrate interactions.

Positions 5 and 9 are marginally unconserved in start TPRs but remain as UMDs in the other two repeat architectures. Despite the large transitions, position 12 is a UMD in middle TPRs. Position 13 is the most missense enriched position in start TPRs and remains missense depleted in middle and end TPRs.

Finally, like CMD site 24, position 23 classifies as a UMD site throughout all three repeat architectures. Located on the convex surface, most substrate binding occurs on the concave surface. It is yet to be uncover its role with substrate interaction, but this site may have a structural role.

Table 3.5: Differences in $PR_{Shenkin}$ and scaled MES between start and middle TPRs. Table is sorted in descending order of scaled MES.
### 3.4 Results and Discussion

<table>
<thead>
<tr>
<th>Alignment Column</th>
<th>( \Delta \text{PR}_{\text{Shenkin}} )</th>
<th>( \Delta \text{Scaled MES} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>64.7</td>
<td>0.8</td>
</tr>
<tr>
<td>15</td>
<td>5.9</td>
<td>0.8</td>
</tr>
<tr>
<td>34</td>
<td>17.6</td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>58.8</td>
<td>0.6</td>
</tr>
<tr>
<td>13</td>
<td>5.9</td>
<td>0.6</td>
</tr>
<tr>
<td>14</td>
<td>17.6</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>11.8</td>
<td>0.5</td>
</tr>
<tr>
<td>17</td>
<td>44.1</td>
<td>0.4</td>
</tr>
<tr>
<td>12</td>
<td>38.2</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>26.5</td>
<td>0.4</td>
</tr>
<tr>
<td>26</td>
<td>11.8</td>
<td>0.4</td>
</tr>
<tr>
<td>29</td>
<td>2.9</td>
<td>0.4</td>
</tr>
<tr>
<td>30</td>
<td>26.5</td>
<td>0.3</td>
</tr>
<tr>
<td>21</td>
<td>5.9</td>
<td>0.3</td>
</tr>
<tr>
<td>19</td>
<td>14.7</td>
<td>0.3</td>
</tr>
<tr>
<td>25</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>33</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>23.5</td>
<td>0.2</td>
</tr>
<tr>
<td>16</td>
<td>23.5</td>
<td>0.2</td>
</tr>
<tr>
<td>24</td>
<td>8.8</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>8.8</td>
<td>0.2</td>
</tr>
<tr>
<td>22</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>27</td>
<td>5.9</td>
<td>0.2</td>
</tr>
<tr>
<td>7</td>
<td>17.6</td>
<td>0.2</td>
</tr>
<tr>
<td>8</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>28</td>
<td>11.8</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>2.9</td>
<td>0.1</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>32</td>
<td>2.9</td>
<td>0.1</td>
</tr>
<tr>
<td>11</td>
<td>5.9</td>
<td>0.1</td>
</tr>
<tr>
<td>18</td>
<td>20.6</td>
<td>0.1</td>
</tr>
<tr>
<td>9</td>
<td>23.5</td>
<td>0.0</td>
</tr>
<tr>
<td>31</td>
<td>5.9</td>
<td>0.0</td>
</tr>
<tr>
<td>23</td>
<td>8.8</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Alignment Column:** High-occupancy alignment columns in the full-length MSA. \( \Delta \text{PR}_{\text{Shenkin}} \): Absolute differences in \( \text{PR}_{\text{Shenkin}} \) between the different repeat architectures. \( \Delta \text{Scaled MES} \): Absolute differences in MES that has been scaled.
3.4.4 Structural analysis of evolutionary Un/Conserved and Missense Depleted positions

ProIntVar retrieved all structures solved for human TPRs in the MSA of Figure 3.5, from the PDBe. VarAlign then collated 276 TPRs from 41 TPR-containing proteins in *Homo sapiens* across 162 structures.

The intra- TPR contact matrix in Figure 3.8A represents all 276 TPRs, highlighting how Helix A and B interact. The inter- TPR contact matrix in Figure 3.8C is made up of four repeat architecture combinations: start-end TPR contacts (9/213; 4%) for 2-TPR-Domains; start-middle (45/213; 21%), middle-middle (114/213; 54%), and middle-end (44/213; 21%) for 3-TPR-Domains and longer. For any interacting pair of TPRs, Figure 3.8C shows that TPR 1 interacts with the full length of Helix $A^2$. Helix $B^1$ interacting with Helix $A^2$ the contact matrix of Figure 3.8C, shows anti-parallel helix packing patterns like intra- TPR contacts in Figure 3.8A. Contrasting this, Helix $A^1$ and Helix $A^2$ interactions Figure 3.8C display parallel helix-helix packing patterns. Although much rarer, Helix $B^1$ and Helix $B^2$ have been observed to interact in 79/213 (37%) TPR pairs. This helical interaction is not as frequently observed, indicating that it is dependent on how the TPR-Domain folds.

RAM filtered out residue pair contacts observed in less than 100/199 EoCs (50%) for the intra-TPR contact matrix and these were listed in Table 3.6. To understand the network of interactions involved with the intra-TPR structural fold, positions that meet the average CN (CN > 2) were visualised in the structure of Figure 5.8B.
Table 3.6: Residue pairs involved with intra-TPR contacts and supporting EoCs. The contact pairs that are observed in the contact matrix of Figure 3.8A. All interactions listed here are residue pairs with more than 50% (100/199) EoCs.

<table>
<thead>
<tr>
<th>X-Axis</th>
<th>Y-Axis</th>
<th>EoCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>11</td>
<td>199</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>195</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>194</td>
</tr>
<tr>
<td>19</td>
<td>11</td>
<td>193</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
<td>193</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>192</td>
</tr>
<tr>
<td>23</td>
<td>7</td>
<td>192</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
<td>192</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>188</td>
</tr>
<tr>
<td>23</td>
<td>11</td>
<td>187</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>182</td>
</tr>
<tr>
<td>26</td>
<td>4</td>
<td>176</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>167</td>
</tr>
<tr>
<td>27</td>
<td>5</td>
<td>166</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>157</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>152</td>
</tr>
<tr>
<td>30</td>
<td>4</td>
<td>147</td>
</tr>
<tr>
<td>24</td>
<td>5</td>
<td>143</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>130</td>
</tr>
<tr>
<td>24</td>
<td>9</td>
<td>101</td>
</tr>
</tbody>
</table>

X-Axis: The alignment column on the x-axis of the contact map in Figure 3.8A. Y-Axis: The alignment column on the y-axis of the same contact map. EoCs: The value representing the evidence of contact.

Visualised in Figure 3.8B, intra-TPR contact analysis identifies CMD site 24 and UMD site 23 with high contact numbers in Figure 3.8A. This is also observed for conserved positions 8, 11, 20, and 27, as well as the unconserved position 1.

RAM filtered out residue pair contacts observed in less than 96/192 EoCs (50%) and these were listed in Table 3.7. Studying the inter-TPR contacts, the forward interactions include 21 and 24 that interact with four positions, whereas positions 12, 17, and 28 interact with three other positions. Contrasting them, backward interactions are dominated by positions 3, 6, 7,
10, and 14. Positions inferred from the sequence and variant analysis are supported by the structural analysis.

Table 3.7: Residue pairs involved with inter-TPR contacts and supporting EoCs. The contact pairs that are observed in the contact matrix of Figure 3.8C. All interactions listed here are residue pairs with more than 50% (96/192) EoCs.

<table>
<thead>
<tr>
<th>X-Axis</th>
<th>Y-Axis</th>
<th>EoCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>7</td>
<td>192</td>
</tr>
<tr>
<td>21</td>
<td>10</td>
<td>188</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>186</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>185</td>
</tr>
<tr>
<td>17</td>
<td>10</td>
<td>181</td>
</tr>
<tr>
<td>21</td>
<td>11</td>
<td>177</td>
</tr>
<tr>
<td>21</td>
<td>7</td>
<td>176</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>176</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>175</td>
</tr>
<tr>
<td>28</td>
<td>4</td>
<td>174</td>
</tr>
<tr>
<td>28</td>
<td>7</td>
<td>174</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>173</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>173</td>
</tr>
<tr>
<td>17</td>
<td>14</td>
<td>172</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>171</td>
</tr>
<tr>
<td>17</td>
<td>13</td>
<td>163</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>160</td>
</tr>
<tr>
<td>24</td>
<td>10</td>
<td>159</td>
</tr>
<tr>
<td>21</td>
<td>14</td>
<td>153</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>146</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>144</td>
</tr>
<tr>
<td>18</td>
<td>14</td>
<td>134</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>132</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>123</td>
</tr>
<tr>
<td>32</td>
<td>4</td>
<td>114</td>
</tr>
<tr>
<td>33</td>
<td>30</td>
<td>103</td>
</tr>
</tbody>
</table>

X-Axis: The alignment column on the x-axis of the contact map in Figure 3.8C. Y-Axis: The alignment column on the y-axis of the same contact map. EoCs: The value representing the evidence of contact.
Firstly, CMD site 24 is observed with a high contact number a second time. Secondly, the two pairs of positions that differ between repeat architectures are important in inter-TPR contacts. Positions 12 and 17, important in start TPRs, have a high number of forward interactions. While the other pair of positions, 3 and 10, are also important for backward interactions in middle and end TPRs. Thirdly, other sites with high contacts, such as positions 7, 21, and 28, are neither significantly enriched nor depleted for missense variants. Fourthly, the same strands true for position 6, which interacts with positions 5, 9, and 12. Thus, position 6 may have high contacts due to the superhelical TPR-Domain fold that permits multiple contacts on the concave surface. Lastly, position 14 is marginally significantly depleted for missense variants ($MES = -0.23$) in end TPRs with backward-interacting positions 17, 18, and 21. Thus, the position may have a structural role end TPRs of a TPR-Domain.

When the intra- and inter-TPR contact numbers were combine to see how all positions within a TPR-Domain interact, positions 4 and 24 stand out here. Position 4 was not classified into any category previously but display a weaker signal in middle TPRs.
3. Tetratricopeptide Repeat Family

3.4 Results and discussion

![Intra-TPR Contact Map](image)

![Inter-TPR Contact Map](image)

![Side view of TPR domain](image)
Figure 3.8: Structural analysis of intra- and inter- TPR contacts. Each contact map is derived from 276 TPRs from 41 TPR-containing proteins, across 162 PDB structures. (A) The intra-TPR contact map shows the residue contacts within each TPRs. (C) The inter- TPR contact map shows the residue contacts between adjacent TPRs. The x-axis to y-axis visualises contacts between alignment positions from the N- to C- terminus direction, and vice versa. The alignment column axes represent equivalent positions in the TPR MSA of Figure 3.5. Cells are shaded according to the EoCs with the colour gradient bar to the right of each contact matrix. Both axes are paired with bar plots, showing the residue pair contacts observed in less than 50% of EoCs in intra- (100/199) and inter- (96/192) TPRs residue contacts. The secondary structure of TPRs is depicted as boxes along both axes. This visualises what residues make up H1, H2, and H3. The lines depict linker peptides between helices. The structure of two adjacent TPRs, numbered 1 and 2, together with the secondary structure on the x and y-axes of the contact maps, orientate the residues involved with interactions (PDB code: 1NA0). The individual positions that meet the average number of contacts, plotted on each bar plot along both contact map axes, have their sidechains visualised and coloured onto CTPR1 for both (B) intra- and (D) inter- CTPR1 contacts.

The RSA for each position across all three repeat architectures were calculated to supplement some results from the contact data earlier. Figure 3.9 shows the median RSA for all three TPR architectures, where the values are listed in Earlier in the variant analysis (Figure 3.7), the pair of positions 12 and 17 are missense depleted and classified as CMD in start TPRs, respectively. Additionally, these positions were important for forward interactions. Unlike positions 3 and 10, these two sites do not change between repeat architectures. Instead, they remain partially exposed throughout the TPR-Domain.
Table 3.8 for each plot. The top three largest changes between start and end TPRs were positions 10, 3, and 6. All sites are highly exposed in start TPRs. Positions 3 and 10 are entirely buried in middle and end TPRs, whereas position 6 reclassifies to partially exposed. Previously, positions 3 and 10 are conserved in middle and end TPRs, as well as missense depleted in end TPRs. Furthermore, in the contact matrices they were recognised to be involved with backward interactions, of which were also the highest CNs. In contrast, position 6 was not a point of interest from the sequence and variant analysis. This site is unconserved on the concave surface, which may have a substrate-binding role and likely not important for the structural fold.
Figure 3.9: Median RSA for each alignment position across start, middle, and end TPRs. The different bars represent the proportion of residues with structural coverage that fall into the different RSA classifications according to the specified thresholds: surface (RSA > 25) partially exposed (5 < RSA < 25) or core (RSA < 5).
Earlier in the variant analysis (Figure 3.7), the pair of positions 12 and 17 are missense depleted and classified as CMD in start TPRs, respectively. Additionally, these positions were important for forward interactions. Unlike positions 3 and 10, these two sites do not change between repeat architectures. Instead, they remain partially exposed throughout the TPR-Domain.

Table 3.8: Median RSA values for all three repeat architectures. The table has been sorted by the absolute differences of the median RSA between start and end TPRs.

<table>
<thead>
<tr>
<th>Alignment Column</th>
<th>Start</th>
<th>Middle</th>
<th>End</th>
<th>Start-End Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>42.9</td>
<td>1.2</td>
<td>0.0</td>
<td>42.9</td>
</tr>
<tr>
<td>3</td>
<td>36.0</td>
<td>0.7</td>
<td>0.0</td>
<td>36.0</td>
</tr>
<tr>
<td>6</td>
<td>43.8</td>
<td>15.6</td>
<td>15.7</td>
<td>28.0</td>
</tr>
<tr>
<td>14</td>
<td>58.6</td>
<td>33.3</td>
<td>35.0</td>
<td>23.6</td>
</tr>
<tr>
<td>7</td>
<td>28.7</td>
<td>7.3</td>
<td>5.1</td>
<td>23.6</td>
</tr>
<tr>
<td>13</td>
<td>68.8</td>
<td>46.9</td>
<td>46.7</td>
<td>22.1</td>
</tr>
<tr>
<td>32</td>
<td>48.2</td>
<td>53.7</td>
<td>60.3</td>
<td>12.1</td>
</tr>
<tr>
<td>11</td>
<td>17.0</td>
<td>11.2</td>
<td>6.6</td>
<td>10.5</td>
</tr>
<tr>
<td>31</td>
<td>24.2</td>
<td>30.0</td>
<td>33.4</td>
<td>9.2</td>
</tr>
<tr>
<td>25</td>
<td>37.6</td>
<td>40.9</td>
<td>45.8</td>
<td>8.3</td>
</tr>
<tr>
<td>15</td>
<td>54.9</td>
<td>59.7</td>
<td>63.1</td>
<td>8.3</td>
</tr>
<tr>
<td>1</td>
<td>10.4</td>
<td>12.2</td>
<td>18.3</td>
<td>8.0</td>
</tr>
<tr>
<td>19</td>
<td>35.0</td>
<td>38.4</td>
<td>42.3</td>
<td>7.3</td>
</tr>
<tr>
<td>33</td>
<td>64.9</td>
<td>62.9</td>
<td>71.0</td>
<td>6.2</td>
</tr>
<tr>
<td>17</td>
<td>23.6</td>
<td>23.8</td>
<td>29.6</td>
<td>6.0</td>
</tr>
<tr>
<td>30</td>
<td>46.7</td>
<td>34.5</td>
<td>41.4</td>
<td>5.2</td>
</tr>
<tr>
<td>2</td>
<td>39.1</td>
<td>30.5</td>
<td>44.4</td>
<td>5.2</td>
</tr>
<tr>
<td>34</td>
<td>27.3</td>
<td>20.7</td>
<td>22.3</td>
<td>5.0</td>
</tr>
<tr>
<td>26</td>
<td>43.2</td>
<td>41.5</td>
<td>39.0</td>
<td>4.2</td>
</tr>
<tr>
<td>28</td>
<td>15.1</td>
<td>13.7</td>
<td>19.2</td>
<td>4.1</td>
</tr>
<tr>
<td>9</td>
<td>32.3</td>
<td>21.5</td>
<td>28.3</td>
<td>4.0</td>
</tr>
<tr>
<td>22</td>
<td>54.7</td>
<td>54.4</td>
<td>51.0</td>
<td>3.7</td>
</tr>
<tr>
<td>23</td>
<td>11.2</td>
<td>10.4</td>
<td>8.1</td>
<td>3.1</td>
</tr>
<tr>
<td>21</td>
<td>14.5</td>
<td>14.3</td>
<td>16.8</td>
<td>2.3</td>
</tr>
<tr>
<td>18</td>
<td>63.7</td>
<td>62.4</td>
<td>65.5</td>
<td>1.8</td>
</tr>
<tr>
<td>16</td>
<td>42.4</td>
<td>38.9</td>
<td>40.8</td>
<td>1.6</td>
</tr>
<tr>
<td>12</td>
<td>23.6</td>
<td>17.3</td>
<td>22.5</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>15.1</td>
<td>9.2</td>
<td>16.2</td>
<td>1.1</td>
</tr>
</tbody>
</table>
Alignment Column: High-occupancy alignment column in the MSA of Figure 3.5. Start: The first TPR of a TPR-Domain. Middle: The bulk of TPRs that contribute to the centre of 3-TPR-Domains or longer. Start: The last TPR of a TPR-Domain. Start-End Difference: Absolute differences between the median RSA values from Start and End columns.

Other sites with variable RSA, which are important from the contact analysis in Figure 3.8E, are 7 and 11. Position 7 is the fifth largest change in RSA ($\Delta RSA = 23.6$), which is entirely exposed in start TPRs, partially exposed in middle TPRs, and marginally buried in end TPRs. This position was not of significance in the sequence and variant analysis, as it alternates between a CME and CMD. Position 11, on the other hand, approximately halves in RSA value from start (17), middle (11.2), and end (6.6) TPRs. But this site remains partially for the first two repeat architectures and is marginally buried in end TPRs.

Lastly, the smallest changes have negligible differences between TPR architectures for positions 8, 20, 27. These three positions, including 24, are heavily buried. Other positions 4, 5, 21, and 23 are partially exposed, with insignificant changes in RSA. Additionally, these sites in Figure 3.8E had a high total number of contact numbers.

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>24</th>
<th>29</th>
<th>27</th>
<th>8</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10.2</td>
<td>9.8</td>
<td>9.3</td>
<td>0.9</td>
<td>0.8</td>
<td>70.9</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1.6</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Figure 3.10: Aggregated TPR substrate interactions. (A) Bar plot showing the total counts of the number of substrate interactions.

Figure 3.10 illustrates all positions that interact with substrate across the 174 TPRs. Substrate interactions occur at every alignment position. This first observation alone contradicts some of what is established in the literature. Structural reviews by Zeytuni and Zarivach (2012) and Perez-Riba and Itzhaki (2019) both discuss key insights into the binding modes of TPR-Domains. However, both fail to discuss how TPR-Domains can organise multi-protein complexes and are capable of convex binding as well. Naturally, the PDBe is dominated by short peptide binding to the concave surface (Perez-Riba and Itzhaki, 2019). However, TPRs can also bind on the convex surface, including the linker regions between the helices. This is especially true for large multi-protein complexes. Examples include x, y, z proteins.

Thresholding at an average number of 40 interactions, substrate interactions are dominated at positions 2, 5, 6, 9, 12, 13 of Helix A. The sequence and variant analysis has identified UMD positions 5, 9, 12, 13, 23, and 34, as potential substrate-binding positions. All but position 23 are consistent with sites enriched for substrate binding. These observations agree with the work of Magliery and Regan (2005), who highlighted that hypervariable
positions in the TPRs are involved in substrate binding. The variant analysis supports majority of hypervariable sites being enriched for substrate binding. From the sites discussed here, Magliery and Regan (2005) have additionally identified positions 2, 6, and 33 to be the other diverse positions that are involved in substrate interactions. Although positions 2 and 6 bind to far more substrates than 33, all three positions are classed as UME in middle TPRs. Thus, may not be binding-specificity positions. Instead, they may be useful for binding affinity by interacting with the backbone of substrates. Since TPR-Domains have a high surface area to volume ratio, it may simply be that these sites are near binding-specificity residues.

On the convex surface, UMD site 23 does not seem to be involved in many substrate interactions. It remains to be seen whether TPRs use this site for substrate binding, but this is unlikely since the analysis supports its new structural role resolved earlier with the contact data. Position 34 however did not have as many substrate interactions as the other UMD sites. This may be because a larger proportion of substrates in the structural data are short peptides. Thus, currently lacking structural data for this site and position 23.

3.4.5 Contrasting with pathogenic variants

Positions in the TPR were investigate further for clinically relevant ClinVar (Landrum et al., 2018) variants that have been experimentally characterised as pathogenic. All pathogenic variants from ClinVar are listed in Table 3.9. Highly conserved or missense depleted positions have been demonstrated to have higher proportions of ‘pathogenic’ or ‘likely pathogenic’ ClinVar variants, termed pathogenic, and vice versa for unconserved or missense enriched sites (MacGowan et al., 2017). Applying the same approach proved difficult to significantly reproduce the reciprocal observations from the variant analysis earlier. However, with a strict PR_{Shenkin} threshold for all TPRs, conserved sites (<20%) had higher rates of pathogenic variants (Fisher OR = 0.97, p = 0.0446, 14 variants) than unconserved sites (>80%, 7 variants). Despite this, a small dataset could not confidently describe the general properties of the TPR family.

Table 3.9: ClinVar variants in TPRs and the variant types.
### ClinVar Variant Types

<table>
<thead>
<tr>
<th>ClinVar Variant Type</th>
<th>Number of Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertain_significance</td>
<td>193</td>
</tr>
<tr>
<td>Pathogenic</td>
<td>35</td>
</tr>
<tr>
<td>Conflicting_interpretations_of_pathogenicity</td>
<td>15</td>
</tr>
<tr>
<td>Likely_pathogenic</td>
<td>15</td>
</tr>
<tr>
<td>Benign</td>
<td>15</td>
</tr>
<tr>
<td>Likely_benign</td>
<td>14</td>
</tr>
<tr>
<td>not_provided</td>
<td>11</td>
</tr>
<tr>
<td>Benign/Likely_benign</td>
<td>10</td>
</tr>
<tr>
<td>Pathogenic/Likely_pathogenic</td>
<td>3</td>
</tr>
</tbody>
</table>

ClinVar Variant Types: The classifications assigned to the variants, if available. Number of Variants: How many variants are classified into the different variant types.

Viewing the distribution of pathogenic variants, Figure 3.11 shows the distribution of variants in the alignment. The CMD sites 4 and 24 have very few positions compared to other dominant ones, like 6 and 20. While not significantly depleted for missense variants, evolutionary conserved positions 7 and 20 have one of the highest numbers of pathogenic variants, indicating pathogenic variants tend to reside at those sites. Unusually, UME sites 2 and 6 have many pathogenic variants. These positions are not interpreted as structurally important, apart from being enriched for substrate interactions. These observations display a near identical distribution of pathogenic variants studied with the Intensification tool by (Chen et al., 2017).
Figure 3.11: ClinVar variants in TPRs. Bar plot of ClinVar variants aggregated at alignment columns of the MSA in Figure 3.5.

It is difficult to interpret any meaningful insights here, since Figure 3.11 shows the concentration of variants are at some positions that are not known to be important from the structural analysis previously discussed. It should be noted that since pathogenic variants are being studied here across individual TPRs in a tabular fashion, a single variant alone can affect a whole TPR-Domain. This can either affect the overall static fold of the TPR-Domain, the flexibility (Llabrés et al., 2020), or be involved with substrate interaction (Xie et al., 2018).

Acknowledging that TPRs have evolved differently in repeat architecture arrangements and accommodating for various substrates, a pathogenic variant at one TPR-Domain may have a different effect in another (Anna et al., 2021). Despite the modular nature of TPRs, the location and effects of pathogenic variants may be unique to specific TPR-containing proteins and the mechanisms by which they function in a system.

3.5 Conclusions
3.5.1 Key insights
MaxHMM updated the set of TPR annotations in the human proteome making up intact TPR-Domains. The reannotation of TPRs have permitted analyses like subfamily analysis: dissecting TPRs according to their PRA has unveiled new conservation profiles of the TPR family, where positions 3 and 10, and 12 and 17 are conserved in start and end
TPRs, respectively; clustering TPRs by sequence similarity identified three large and distinct subfamilies that make up 3-TPR-Domains involved with Hsp70 and 90 binding.

The variant analysis identified new positions outside of the conserved set of positions characteristic of the TPR family. The established conserved positions 8 and 20 were not the most constrained in the human population (Figure 3.7). Instead, it was both conserved and unconserved missense depleted (CMD/UMD) sites 24 and 23, respectively, that were conserved for large, hydrophobic and aromatic amino acids (Figure 3.6E). Extending the insights from the new conservation profiles observed in a TPR-Domain, positions 3 and 10 are missense depleted and CMD, respectively, in end TPRs. Whereas positions 12 and 17 are missense depleted and CMD, respectively, in start TPRs. The most comprehensive structural analysis of human TPRs to date was conducted here, visualising all residue pair contacts within and between TPRs of a TPR-Domain. The intra- and inter-repeat contact data dissects the structural roles of each position within the TPR-Domain. Finally, the structural data validated the CMD site 24 and UMD site 23 with high contacts critical for the TPR-Domain fold. These two positions occupied the major clefts between the same helix type of adjacent TPRs. In constrast, the rest of the UMD sites are enriched for substrate interactions and may be responsible for substrate binding specificity.
4.1 Preface

This chapter is based on joint work with Javier S. Utgés, Noah J. M. Dietrich, and Stuart A. MacGowan. When developing RAM with the TPR family, I co-supervised Noah as a summer intern, and he started the work by applying the components of RAM to the Ankyrin repeat family. These components, originally written by me, include gathering repeat annotations, PRA analysis, and building MSAs. Javier, whom I co-supervised as a Master’s student, continued Noah’s work by improving the first three components of RAM (Steps 1–3 in Figure 2.1) and adding new modules to them. Javier completed the entire analysis on the Ankyrin repeat family, and the work was published under the title “Ankyrin repeats in context with human population variation” (Utgés et al., 2021).

As first developed with the Ankyrin repeat family, Javier redefined the surfaces of Ankyrin-Domains into three surfaces. These new findings contrast with the long-established concave and convex surfaces. The approach implemented here was a quantitative approach to
define the surfaces, where positions with the highest RSA values represented the boundaries of each surface. Surfaces have often been defined by approximating which positions make up the different surfaces from sampled protein structures, instead of a quantitative metric. Javier redefined new surfaces using median RSA values across all structures for sequences in the MSA.

Another addition to the structural analysis in RAM was his assignment of the correct DSSP definition of the SSEs observed across all available structures. Consequently, Javier further classified beta-turns into Asx motifs in Ankyrin repeats. These new additions were assimilated into RAM as permanent analytical modules.

4.2 Introduction

Breeden and Nasmyth (1987) reported a repeating stretch of approximately 33 amino acid long sequences in the Swi6p and Cdc10p yeast cell-cycle regulators, Notch and LIN-12 developmental regulators in *D. melanogaster* and *C. elegans*. Not long afterwards, similar repeat sequences were discovered in the human erythrocyte protein ankyrin, after which they were first named (Lux et al., 1990). Since then, Ankyrin repeats have been discovered in many other functionally diverse proteins (Reviewed in Bork, 1993). These include GA-binding protein subunit beta-1 (GABP-1) is a transcription factor that interacts with purine rich repeats (GA repeats); the ANKRD31 functions in cell cycle regulation, where it is required for DNA double-strand breaks (DSBs) that initiate meiotic recombination (Papanikos et al., 2019); cytoskeletal integrity; ion transport, such as the Transient Receptor Potential Vanilloid receptors that modulate intracellular calcium levels (Phelps et al., 2007); Neurogenic locus notch homolog protein 1 (Notch) is a transmembrane receptor for the highly conserved signalling pathway that regulates transcription for crucial for development (Brütsch et al., 2010).
Like TPRs and Armadillo repeat families, Ankyrin-containing proteins are prominent in eukaryotes. The presence of so many functionally diverse Ankyrin-containing proteins could be the result of domain shuffling (Bork, 1993). However, comparing the three repeat families, the number of species having one or more Ankyrin-containing proteins is least in prokaryotes, and quite rare in archaea (Jernigan and Bordenstein, 2015, Kamel et al., 2021). However, it is prominent in protozoa (Kamel et al., 2021). This taxonomic distribution may be the result of horizontal gene transfers (Bork, 1993).

Ankyrin repeats are 33 amino acids long with a right-handed fold, shown in Figure 4.1B, that shapes into unique and conserved anti-parallel α-helices, as well as a β-hairpin/loop in some known structures (Gorina and Pavletich, 1996). Adjacent Ankyrin repeats are virtually parallel to each other with a 2–3° counter-clockwise rotation. As a result, Ankyrin-Domains in Figure 4.1C form an extended helical bundle with a small left-handed twist (Kobe and Kajava, 2000). Additionally, in some cases the β-hairpin/loop can form a continuous anti-parallel β-sheet. The β-hairpin is perpendicular to the α-helices that form an L-shaped structure (Gorina and Pavletich, 1996). Helix 1 (inner) is seven residues long that spans positions 5–12, whereas the Helix 2 (outer) is nine residues long that extends from positions 15–24.
Figure 4.1: Ankyrin repeat sequence and structure. (A) Sequence logo of the Ankyrin MSA summarised in Figure 4.3, generated with WebLogo on the online webserver tool (Crooks et al., 2004). The y-axis represents the probability of observing an amino acid at any position in the alignment. (B) Ribbon representation of an Ankyrin repeat, where the α-helices are in red and the coil in blue. The repeat is numbered according to the sequence logo. (C) Surface and ribbon representation of an Ankyrin-Domain from the human Gankyrin protein (O75832). Adjacent repeats are highlighted with alternating colours of orange and blue. Start and end repeats are coloured in purple and green, respectively. PDB code: 1UOH (Krzywda et al., 2004).

The sequence logo in Figure 4.1A highlights the key residues important for the fold of the Ankyrin repeat. The most prevalent is the TPLH motif at positions 4 through 7. Proline residues at position 5, which resides at the beginning of Helix 1 in Figure 4.1B, promotes the tight turn to permit the L-shaped fold of the Ankyrin repeat. Additionally, threonine and
histidine residues at position 4 and 7 stabilise the β-hairpin/loop regions, respectively, with hydrogen bonds. Lastly, glycine residues are conserved at positions 13 and 25 to promote the tight turning of the anti-parallel α-helices and the loop that extends underneath the helical bundle towards the β-hairpin/loop region, respectively (reviewed in Mosavi et al., 2004).

The helical bundle is supported by a network of interactions, where positions 9 and 18 are involved with intra-repeat contacts. In contrast, positions 8, 10, 17, 20, and 22 are believed to be involved with inter-repeat contacts that all make up the interactions stabilising the Ankyrin-Domain. Lastly, a hydrophobic core runs through the complete length of the Ankyrin-Domain comprised of positions 6 and 21 (reviewed in Mosavi et al., 2004).

Ankyrin repeats have a unique structural fold that differs from other all-helical repeat families in Section 1.4.3. The other solenoid-repeat families covered in this thesis have concave surfaces contributed by flat helices that twist with every successive repeat. However, Ankyrin-Domains adopt a distinctive L-shaped fold that permits a semi-enclosed surface to bind to substrates. The binding surface is contributed by Helix 1 and the β-hairpin/loop regions. Canonical positions on the concave surface are highly diverse to accommodate a wide repertoire of protein substrate interactions (Mosavi et al., 2002, Mosavi et al., 2004). In fact, some Ankryin-containing proteins are also capable of binding to lipids and sugars (Islam et al., 2018). Specifically, Magliery and Regan (2005) identify positions 2, 3, 5, 13, 14, 17, 32, and 33 as specificity-determining sites by their sequence variation using statistical free energy analysis. Alternatively, Mosavi et al. (2004) uses Information Content generated from ALPRO (Schneider and Stephens, 1990) to identify positions 1, 3, 11, 12, 15, 25, and 33 that score less than 1 bit.

4.3 Methods
4.3.1 Gathering repeat annotations
Inspecting all InterPro Entries that annotate Ankyrin repeats across all Biological Entities, IPR002110 best represents them. Databases SMART (SM00248), ProSite (PS0088),
PRINTS (PR01415) and PFAM (PF13606, PF00023, PF13637, PF13857, PF12796) contribute to this InterPro Entry. Pfam entries PF13637, PF13857, and PF12796 were not included in the analysis here, since annotate two or more repeats. The total number of annotations across all four member databases and Swiss-Prot, with keyword "ANK repeat [KW-0040]", are presented in Figure 4.2B and again as a horizontal bar to the left of the UpSet plot in Figure 4.2A.

Figure 4.2: Composition of the Ankyrin repeat annotation dataset. (A) An UpSet plot visualising how Pfam (PF13606, PF00023), PRINTS (PR01415), PROSITE (PS50088), SMART (SM00248), and Uniprot from Swiss-Prot ("ANK") databases annotate Ankyrin repeats in 1294 Ankyrin-containing proteins across 133 organisms. (B) Vertical bar plot of the total number of Ankyrin annotations from different databases. The plot is identical to the horizontal bar plot to the left of the UpSet plot in panel A. (C) The number of annotations from each database contributing to the final non-redundant set of Ankyrin annotations. Figure 3 from Utgés et al. (2021).
As explained in Section 2.3.1.2, all the databases were inspected for the proportion of a canonical Ankyrin repeat that they annotate. The PRINTS database annotated a small fraction of a canonical Ankyrin repeat and were deemed unfit for this analysis. Conveniently, when comparing all the annotations, the UpSet plot in Figure 4.2 shows how the other databases, like Swiss-Prot and SMART, have higher coverage than that of PRINTS, which has no unique annotations itself. Therefore, a non-redundant merge of all annotations was carried out in this order: PROSITE, SMART, Swiss-Prot, and Pfam. Summarised in Figure 4.2, the final dataset of 7,407 annotations consists of 4,109 PROSITE, 2,313 SMART, 972 Swiss-Prot, and 10 Pfam (PF00023 and PF13606).

4.3.2 Multiple sequence alignment
Several alignment methods were applied to Ankyrin repeats to assess their performance. Sequence alignment methods included ClustalΩ, HMMER, T-Coffee, pairwise alignment with Multalin (AMPS), Muscle. STAMP, a structure-based alignment method, was also carried out. As for the Armadillo repeat family (Chapter 5), all methods performed poorly. These alignment methods introduced insertions within SSE of Ankyrin repeats. Furthermore, ranges of columns had misaligned residues that could not recapitulate the conserved positions of Ankyrin repeats. The difficulty in aligning Ankyrin repeats led to the development of the new alignment algorithm described in Section 2.3.3.1. This recursive algorithm successfully aligned approximately 98% of sequences and the unaligned sequences (2%) were realigned to the resulting MSA with a profile–profile alignment, producing the final MSA in Figure 4.3.
4. Ankyrin Repeat Family

4.3 Methods

Figure 4.3: Overview of the Ankyrin MSA. The MSA comprises 7404 Ankyrin repeats. Sequences were sorted by an average distance tree calculated Jalview with the BLOSUM62 substitution matrix. The MSA was visualised in the Overview Window of Jalview with the ClustalX colour scheme: hydrophobic residues, in blue; polar, in green; proline, in yellow; glycine, in orange; unconserved, in white; gaps, in grey. The annotation ladder numbers high-occupancy (>50%) alignment columns. The SS annotation track at the bottom of the figure represents SSEs obtained from Figure 4.12, where the red boxes represent helices. The unbroken lines represent linker regions without SSEs, and dashed lines represent low-occupancy columns. Figure 4 from Utgé et al. (2021).

4.3.3 ProIntVar

VarAlign mapped 419 sequences, using ProIntVar, to 209 X-ray crystallography structures from the PDB. ProteoFAV (MacGowan et al., 2017) retrieved the real-space R value (RSR), RSR-Z scores, and real-space correlation coefficient (RSCC) quality metrics (Kleywegt et al., 2004) from the validation reports in the PDBe. Residues with RSCC < 0.85 and RSRZ > 2 were filtered out, leaving 383/419 Ankyrin repeats across 73/80 proteins from 176/209 structures. The final dataset had an average RSRZ per residue (11,186/13,059) of 0.11 and a mean RSCC of 0.95.

4.3.4 Structural analysis

The structural analysis follows the same methodology introduced in Section 2.3.8, but with some exceptions. Instead of only using absolute EoC values in contact matrices, they are normalised by the coverage of every residue pair observed within a structure; termed structural
coverage. Equation 6 calculates the total number of contacts per residues ($C_i$), where $C_{ij}$ is the incidence of EoCs for a pair of residues at positions $i$ and $j$ within $K=33$ positions in the Ankyrin MSA. Similarly, Equation 7 calculates the total structural coverage ($O_i$) by adding the absolute frequency of positions $i$ and $j$ being present within a repeat.

$$C_i = \sum_{j=1}^{K} C_{i,j} \quad (6)$$

$$O_i = \sum_{j} O_{i,j} \quad (7)$$

Lastly, Equations 8 and 9 sum the contacts and coverage, respectively, across the 33 positions in the Ankyrin repeat.

$$C_t = \sum_{i} C_i \quad (8)$$

$$O_t = \sum_{i} O_i \quad (9)$$

Finally, the contact number analysis adopts a similar strategy to MES in the variant analysis (Section 2.3.7.2). The calculation involves applying the natural logarithm of an OR for contact number values, which are observed at each position in the intra-Ankyrin contact matrix that have been normalised by structural coverage. This transformation is termed contact enrichment score (CES).

PPIs were initially normalised by the structural coverage of structures observed to interact with a protein substrate. PPIs are then measured in the same manner as MES, and the contact number analysis here is termed PPI enrichment score (PPIES).
Figure 4.4 identifies positions 4, 5, and 6, which contribute to the TPLH motif of Ankyrin repeats as the most conserved sites with $N_{\text{Shenkin}} < 25$. This set also includes sites 9, 21, and 22 with sidechains in Figure 4.1, facing an adjacent repeat in the C- to N-terminus direction. Furthermore, position 13 is another conserved site located in the linker region between Helix 1 and 2. Contrasting these positions are ten divergent sites that score $N_{\text{Shenkin}} > 75$. Most of the divergent sites on the concave surface area are involved with the bulk of substrate interactions.

![Image](image.png)

Figure 4.4: Normalised Shenkin conservation profile. Sequence conservation profile of the Ankyrin MSA from Figure 4.3 for high occupancy alignment columns with 50% or more residue occupancy. Positions were categorised based on the conservation quartile they belong to: $N_{\text{Shenkin}} \leq 25$, in blue; $25 \leq N_{\text{Shenkin}} \leq 50$, in green; $50 \leq N_{\text{Shenkin}} \leq 5$, in yellow; $75 \leq N_{\text{Shenkin}} \leq 100$, in red. Figure 6A from Utgés et al. (2021).

4.4.2 Variant analysis

VarAlign retrieved and mapped 21,338 missense variants to 1,435 human Ankyrin repeat sequences in the MSA. Next, RAM calculated the MES for each position, and the values were plotted against $N_{\text{Shenkin}}$ in Figure 4.5. Thresholds laid out in Section 2.3.7.2 of the Methods chapter classified four sets of positions in Figure 4.5: CMD sites 6, 9, 13, 21, and 22; CME sites 4 and 5; UMD sites 1, 3, 8, 33; UME sites 11, 12, 15, 24, 30, 31.
Figure 4.5: Missense enrichment plot. Scatter plot showing the relationship between MES and $N_{Shenkin}$ (Figure 4.4). The numbers at each data point refer to the column positions in the Ankyrin MSA of Figure 4.4. Each datapoint that resides above (enriched) or below (depleted) an MES of 0, as well as reside in the lower (25%) and upper (75%) quartile $N_{Shenkin}$, have been classified into four categories: Conserved and Missense Depleted (CMD) in blue, Conserved and Missense Enriched (CME) in green, Unconserved and Missense Depleted (UMD) in red, Unconserved and Missense Enriched (UME) in yellow. Unclassified positions were in grey. Error bars indicate 95% CI. Figure 9A from Utgéš et al. (2021).

4.4.2.1 Conserved and missense depleted positions

Figure 4.5 identifies positions 6, 9, 13, 21, and 22 as CMD sites. With reference to the structure of an Ankyrin repeat in Figure 4.1, all sites, apart from 13, are buried. Furthermore, CMD sites are characterised by hydrophobic amino acids observed in the MSA of Figure 4.3. Position 13 instead is conserved for glycine and is therefore constrained in the population to preserve the tight-turning region. Other missense depleted sites in Figure 4.5 that did not meet the conservation threshold include positions 7 and 32. These two positions reside in the linker regions before Helix 1 and after Helix 2 that make up the start and end positions of an Ankyrin repeat, respectively. Since these two positions may be necessary for the stability of that region,
they will be included with the CMD set of sites as well. All six CMD sites are likely to be important for the structural fold.

4.4.2.2 Unconserved and missense depleted positions

It is well known that solenoid proteins have positions specialised in substrate interactions that tend to be unconserved or diverse in amino acid conservation (Mosavi et al., 2002, Magliery and Regan, 2005). The concave surface of Ankyrin-Domains is known to display unconserved sites that accommodate the diversity of substrates they can interact with. From the unconserved sites in Figure 4.5, the variant analysis identified unconserved positions depleted for missense variants relative to other unconserved sites. All four UMD sites 1, 2, 8, and 33 are located on the concave surface, indicating that they are probably responsible for substrate-binding specificity.

4.4.3 Structural analysis

A collection of 176 three-dimensional structures were retrieved via ProIntVar, representing a total of 383 different Ankyrin repeats from the PDBe. This structural data was aggregated, as explained in Section 4.3.3, to study and support the sequence and variant analysis earlier with intra-Ankyrin contact analysis.

4.4.4 Intra-Ankyrin contacts

The contact map in Figure 4.6 shows the interactions observed in intra-Ankyrin contacts. The local interactions displayed as single strips of contacts and blocks of contacts highlight the typical patterns of linker regions and the two helices, respectively.

All CMD sites identified earlier in Figure 4.5, apart from position 13, were enriched for intra-Ankyrin contacts in Figure 4.6B. These sites are important for the fold of Ankyrin repeats displayed in the structure of Figure 4.6C. This observation explains why these CMD sites are conserved across homologues and constrained in the human population. In addition, inter-Ankyrin contact analysis did not show any relation between enrichment of inter-repeat contacts and sequence and variant classifications earlier. Position 13, on the other hand, is conserved
for glycine residues. Although the structural analysis alone does not support its CMD classification, it is instead likely important for the turn between the two helices.
Figure 4.6: Intra-Ankyrin contact analysis. (A) Contact map of intra-Ankyrin residue interactions within each Ankyrin repeat. Cells are coloured according to the colour gradient, that represents the proportion of contacts observed between residue pairs. SSE helix alpha-
helices are depicted as red boxes. (B) Intra- Ankyrin enrichment plot. The y-axis named Intra-ANK contacts enrichment score refers to the CES (See Section 4.3.4). Data points are coloured by the sequence and variant analysis classifications assigned to them. Error bars represent 95% CI of CES. (C) Intra- Ankyrin contacts (orange lines) between Helix 1 and 2, and (D) contacts between positions 1–7 and 28-32. PDB code: 5MA3. Figure 10 from Utgés et al. (2021).

4.4.5 Median RSA reclassifies established surfaces of Ankyrin-Domain

To identify positions enriched for substrate interactions, the initial quantitative determination of which residues contribute to which surface is key. Visualised in the structures in Figure 4.8, the literature has previously established concave (positions 32–12) and convex (positions 13–31) surfaces (Wang et al., 2014). From the structural data aggregated on a column basis, Figure 4.7 shows the median RSA in the MSA. Positions with large median RSA values, together with guidance of an Ankyrin repeat structure, were used to define the borders of surfaces on Ankyrin-Domains.

![Figure 4.7: Median residue surface accessibility. Bar plot of median RSA. Error bars represent 95% CI of median RSA. Bar plot of the median RSA for each alignment position. Positions have been grouped into different RSA classifications according to these specified thresholds: surface (RSA > 25) partially exposed (5 < RSA < 25) or core (RSA < 5). Figure 6C from Utgés et al. (2021).](image)

First, positions classified as buried in Figure 4.7 were grouped together in Table 4.1, and the rest of the positions were distributed to different surfaces. Therefore, when the ranges of values defining surfaces are described here, those buried sites from Table 4.1 are excluded.
All ranges of values are in the N- to C-terminus direction; for example, the range 31–8 will include 31, 32, 33, 1, 2, 3, and 8.

Figure 4.8: Redefined surfaces of Ankyrin-Domains. (A, B) The original and (C, D) the new classifications of Ankyrin-Domain surfaces. (A, C) Surface representation of the last 12 Ankyrin repeats from the D34 ANK1 Ankyrin-23-Domain. Surface representations are coloured by the key underneath each structure. (B, D) Individual Ankyrin repeats highlighting the surfaces of panel A and C, respectively. PDB code: 1N11 (Michaely et al., 2002). Figure 8 from Utgé et al. (2021).

Positions 32–12 in Figure 4.8B originally defined the concave surface in Figure 4.8A. Studying those literature-based concave positions in Figure 4.7, the RSA value at position 12 is a local maximum and positions 32 and structurally neighbouring positions 33 and 1 (local maxima) are in an Ankyrin-Domain. Therefore, the concave surface, visualised in Figure 4.8C, will be defined in Table 4.1 between positions 32–12 of Figure 4.8D.
Table 4.1: Classifications of positions with median RSA. Positions were classified as buried, concave, convex, or basal using median RSA values from Figure 4.7. This was carried out with the support of an Ankyrin structure in Figure 4.1.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Canonical positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buried</td>
<td>4, 5, 6, 7, 9, 10, 17, 18, 21</td>
</tr>
<tr>
<td>Concave</td>
<td>1, 2, 3, 8, 11, 12, 32, 33</td>
</tr>
<tr>
<td>Convex</td>
<td>13, 14, 15, 16, 19, 20, 22, 23, 24</td>
</tr>
<tr>
<td>Basal</td>
<td>25, 26, 27, 28, 29, 30, 31</td>
</tr>
</tbody>
</table>

Classification: The region assigned to an Ankyrin-Domain. Canonical positions: The position in the alignment of Figure 4.3.

Next, the concave surface in Figure 4.8A was originally defined between positions 13 and 31 (Figure 4.8B). Viewing those positions in Figure 4.7, position 23 is a global maximum interrupting that range. Inspecting positions 23–25 in Figure 4.1B, these sites define a clear ridge between Helix 2 and the linker region that runs underneath the two helices of the Ankyrin repeat. Those three sites were used to split the originally defined convex surface into two surfaces: convex and basal. The convex surface (Figure 4.8C) is here classified between 13–24 in Figure 4.8D, while the new basal surface is characterised between 25–31.

4.4.6 New surface classifications reflect depletion of missense variants.

The relationship between the sequence and variant analysis, from Section 5.5.2, was compared with the RSA properties. Buried residues, from Table 4.1, were depleted for missense variants ($MES = -0.10, p < 1.9 \times 10^{-7}$) relative to residues present on the surface.

Furthermore, the sequence and variant classifications were revisited in the context of the new surface classifications in Figure 4.9. Canonical positions that make up the concave surface are significantly depleted for missense variants ($MES = -0.08, p < 4.4 \times 10^{-4}$), relative to other surfaces. Conversely, the basal surface was significantly enriched for missense variants ($MES = 0.09, p < 6.2 \times 10^{-6}$), unlike the convex surface that was neutral in comparison with the...
other two surfaces. Lastly, the basal surface again displayed significantly high MES values compared to the convex ($MES = 0.08, p < 8.8 \times 10^{-4}$).

Figure 4.9: MES on Ankyrin-Domain surfaces. Surface representation of the last 12 Ankyrin repeats from the D34 ANK1 Ankyrin-23-Domain. Four different views of the Ankyrin-Domain, visualising all surfaces and the three new surface classifications independently. The surfaces are coloured by the colour gradient on the right, between the highest and lowest MES values observed in the MSA. The surfaces not highlighted are coloured in grey. PDB code: 1N11 (Michaely et al., 2002). Figure 9B from Utgés et al. (2021).

4.4.7 Protein substrate interactions

The literature has established that the concave surface has high sequence variability to enable it to bind to a large repertoire of protein substrates (Magliery and Regan, 2005). There are 142 Ankyrin repeats across 35 Ankyrin-Domains that interact with 63 protein substrates. With the newfound surfaces established for the Ankyrin-Domains, the PPIES was measured, and displayed in Figure 4.10A, to identify positions enriched for substrate interactions.

All positions on the concave surface in Figure 4.10A are enriched for PPIs. Other sites, including the buried CMD site 7, are accessible on the concave surface. Positions 13 and 14 at one edge of the convex surface also have high PPIES. This observation is due to their proximity to the concave surface. The analysis continues by collectively comparing PPIES at each position between the surfaces to which they belong in Figure 4.10B.
4. Ankyrin Repeat Family

4.4 Results and discussion

Figure 4.10: Protein-substrate interactions. Dot plot of PPIES values for each canonical position in the Ankyrin repeat, coloured by the new surface classifications (Table 4.1). Error bars represent 95% CI of PPIES (B) Surface representation of the last 12 Ankyrin repeats from the D34 ANK1 Ankyrin-23-Domain. Four different views of the Ankyrin-Domain, visualising all surfaces and the three new surface classifications independently. The surfaces are coloured by the colour gradient on the right, between the highest and lowest PPIES values observed in the data set. The surfaces not highlighted are coloured in grey. PDB code: 1N11 (Michaely et al., 2002). Figure 11 from Utgés et al. (2021).

Buried residues were significantly depleted for PPIs relative to other surfaces ($PPIES = -1.02$, $p < 10^{-16}$). Concave residues are highly enriched in PPIs
Ankyrin Repeat Family

4.4 Results and discussion

\( PPIES = 1.86, \ p < 10^{-16} \). In contrast, residues on the convex and basal surfaces are relatively depleted for PPIs with \( PPIES = -0.79 \) and \( p < 10^{-16} \), and \( PPIES = -2.19 \) and \( p < 10^{-16} \), respectively. Lastly, convex is relatively more enriched for PPIs than the basal surface.

The PPIES values displayed on the surface of an Armadillo-Domain in Figure 4.10B, these observations confirm that the concave is the main mode of substrate interaction. In addition, convex binding is rarely observed, and the basal surface has little to no role with substrate binding.

The UMD sites on the concave surface are enriched for PPIs \( PPIES = 3.6 \ p < 10^{-16} \). Since these sites are constrained in the human population, they are likely to be functionally important for substrate interaction. Figure 4.11 visualises the UMD sites in the context of Ankyrin-Domains from two different Ankyrin-containing protein families bound to substrates with motifs.

Tankyrase-1 and -2 are closely related poly(ADP-ribose) polymerases that employ all UMD and UME sites on the concave surface of Ankyrin-Domains, in Figure 4.11 A and B, to bind perpendicularly with the RXXPDG motif of USP25 and ARPIN2 proteins, respectively (Sbodio and Chi, 2002). A similar arrangement is observed for the DNA-binding protein RFXANK and Ankyrin repeat family A protein 2 substrates, involved in transcription regulation of RNA polymerase II. These two proteins, and five other known protein substrates, have an PXLPX[I/L] motif that, instead, extends across the entire length of the Ankyrin-Domain in these examples (Nie et al., 2015, Xu et al., 2012) and many other structures.
4.4 Results and discussion

These examples demonstrate how Ankyrin-Domains recruit UMD sites to bind to motifs and similar protein substrates. These observations further support the key findings observed across all protein families, where UMD sites are likely binding-specificity residues (MacGowan et al., 2017).

4.4.8 Secondary Structure classification redefines beta strand

The summarised SSEs for all structures, in Figure 4.12, assign the SSE across the whole repeat family; Helix 1 and Helix 2 reside between positions 5–11 and 15–23, respectively.
Figure 4.12: Aggregated SSE classifications. Stacked bar plot of the summarised SSE assignment across all structures at each position. Each position represents the proportion of each SSE assignment as defined in DSSP: α-helix, 3₁₀-helix, π-helix, β-bridge, β-strand, turn, bend and coil. Figure 6B from Utgés et al. (2021).

Other distinct SSEs are four pairs of adjacent positions that contribute to turns. There are two 5-turns at positions 12–13 and 24–25 that reside at the ends of both α-helices. Conversely, positions 28–29 and 33–3 mimic type I β-turns. Positions 27 and 32 (residue i) here are conserved for asparagine or aspartate with residue occupancy of 44% and 58%, respectively. Accordingly, both have been classified as Asx motifs (de Brevern, 2016) in Figure 4.13: a hydrogen bond is formed between the sidechain CO group of residue i (positions 27 and 32) and the mainchain NH of residue i + 2 (positions 29 and 1), and a hydrogen bond between the mainchain CO group of residue i and the mainchain NH of residue i + 2 (positions 30 and 2). Turn 27–30 was classed as an Asx-β-turn in Figure 4.13A. In contrast, turn 32–2 was classed as a type 1 β-bulge loop with an Asx motif in Figure 4.13B. This turn has an additional
hydrogen bond between the sidechain CO group of residue i + 4 (position 3) with the mainchain NH of residue i (position 32).

Ankyrin repeats that do not have an Asx motif, due to the absence of an Asx residue at positions 27 and 32, will form a simple β-turn instead. This SSE is incapable of folding due to the lack of hydrogen bonds. The conservation of Asx residues at those sites demonstrate the importance of forming the iconic β-turns needed for the fold of the Ankyrin-Domain.
Figure 4.13: Asx motifs in the Ankyrin repeat loop. (A) Asx-β-turn and (B) type I β-bulge loop with an Asx motif. Two numbering systems refer to the canonical Ankyrin positions (e.g. 27, 28, 29, etc.) and the naming convention for the position of residues on the primary sequence level (i, i + 1, etc.). Red and blue rods are oxygen and nitrogen atoms, respectively. PDB code: 5MA3 (Hansen et al., 2017). Ribbon and transparent surface representations of an Ankyrin-Domain with the Asx residues displaying sidechains in both Asx turns. PDB code: 2Y1L
4.5 Pathogenic variants

Acknowledging the new data from the sequence, variant, and structural analysis, pathogenic variants were inspected to understand how they behave in this repeat family. In line with observations in TPRs, Figure 4.14 also shows there are too few variants to confidently identify positions enriched or depleted for pathogenic variants. Furthermore, the distribution in Ankyrin repeats again shows unusual trends that would contradict the evolutionarily conserved or missense depleted sites. Positions of interest include 5, 7, and 13, but they do not coincide with key sites inferred from the sequence, variant, and structural analysis. All three sites are enriched for PPIs and may have an influence on altering binding to substrate partners, but it is difficult to confidently interpret these results.

Figure 4.14: ClinVar variants in the Ankyrin repeat family. Variants considered have these clinical significance values: likely pathogenic, pathogenic, risk factor, drug response. Red line represents the average of all ClinVar variants (2.6).

4.6 Conclusions

4.6.1 Key insights

Conserved and missense depleted sites (CMD) are significantly enriched in intra-repeat contacts that are important for the structural fold of the Ankyrin repeat. Whereas
unconserved and missense depleted sites (UMD) are significantly enriched in in protein-protein interactions (PPIES = 3.6, $p < 10^{-16}$) that may have a binding-specificity role.

The surface of Ankyrin-Domains has been redefined into three distinct surfaces using aggregated RSA data across all available structures at each position. The variant analysis was additionally analysed across multiple positions that contribute to newly classified surfaces. The concave surface of Ankyrin-Domains is depleted for missense variants (MES = -0.08, $p < 4.4 \times 10^{-4}$) and significantly enriched in PPIs (PPIES = 1.86, $p < 10^{-16}$). However, both convex and basal surfaces are not under constraint in the human population with little to no role with substrate interactions.

Finally, collating a large collection of structural data, the summarised SSEs of turns 28-29 and 33-1 in the Ankyrin repeats are defined as Asx motifs, where positions 27 and 32 are conserved for Asx residues.
Chapter 5

Armadillo Repeat Family

5.1 Preface

This chapter is based on joint work with Thibault Rosazza, a first year Wellcome Trust student doing a rotation project under my supervision. Thibault applied the same principles and methods outlined in RAM to the Armadillo repeat family, and compared his analysis with that of other repeat families to understand how to develop RAM further. Thibault was involved in the early stages of the work, and I completed the project.

One of the difficulties encountered in the project was aligning Armadillo sequences using standard alignment methods. Thibault did not use the recursive alignment method described in Section 2.3.3.1, but managed to get a reasonable alignment with the ClustalW alignment method and manual curation. The exact same positions covered in my work were observed in Thibault’s project.
5. Armadillo Repeat Family

5.2 Background

When studying the Armadillo repeat family with the RAM pipeline, aspects of the Armadillo family raised specific issues that RAM was not originally designed to address. The development of new approaches to tackle these problems led to their eventual inclusion in RAM.

In an attempt to align Armadillo repeat sequences, the popular alignment methods produced the same issues, giving poor MSAs with many insertions across the entire alignment. The solution was to build MSAs with the MULTALIGN program in AMPS that implements secondary-structure-based gap penalties. AMPS significantly improved the alignment of repeats, and this method became the foundation for aligning repeat sequences in RAM (Section 2.3.3.2). Although not covered in the thesis, this approach was tested on Ankyrin repeats and TPRs for which AMPS successfully generated high-quality alignments with far fewer insertions than standard alignment methods. Furthermore, a qualitative approach to evaluate the quality of MSAs from a selection of alignment methods was developed and incorporated into RAM.

The variant analysis classified three CME sites (Section 5.5.2), and the RSA analysis (Section 5.5.3.1) classified them as buried. The structural data contradicts the classification assigned to these three positions since they are enriched for missense variants relative to other conserved positions. Accordingly, substitution matrices of all the variants at CME sites were visualised in Section 5.5.2, to characterise the classifications that conflict with the structural analysis. This section explores how these sites were enriched for conservative mutations with similar physicochemical properties observed in the alignment at those positions. Thus, substitution matrices were incorporated into RAM to characterise the variant properties of positions in a repeat family.
5.3 Introduction

Riggleman et al. (1989) first discovered Armadillo repeats in the Drosophila segment polarity gene product Armadillo. The phenotype of the of this homozygous lethal gene defect is that Drosophila larvae resemble an armadillo, thus giving rise to its name. Since then, several Armadillo homologues have been cloned and sequenced and the proteins studied. Examples include the tumour suppressor adenomatous polyposis coli (Peifer et al., 1994), the junctional plaque protein plakoglobin (Franke et al., 1989), and the nucleocytoplasmic transport factor importin-α (Görlich et al., 1994). In addition, Armadillo-containing proteins function in a broad range of cellular processes such as molecular signalling (Cadigan and Peifer, 2009, Gates and Peifer, 2005), cytoskeletal regulation (Gates and Peifer, 2005), cell–cell adhesion (Cadigan and Peifer, 2009), intracellular transport (Manning and Snyder, 2000), protein degradation (Downes et al., 2003), molecular chaperones (Kim et al., 2008), regulation of gene expression during development (Cadigan and Peifer, 2009, Daniel and Reynolds, 1999), and other roles often including multiple combinations of these functions (Tewari et al., 2010).

Armadillo-containing proteins are found throughout evolution (Tewari et al., 2010, Gul et al., 2017, Jernigan and Bordenstein, 2015). Like many repeat families, certain Armadillo-containing protein families have developed specific functions within certain species, dictated by function and evolution. Thus, the repertoire of Armadillo-containing proteins from simple, unicellular, organisms has undergone a considerable expansion with increasing complexity in multicellular organisms, or been lost in particular species lineages (Tewari et al., 2010, Jernigan and Bordenstein, 2015). For example, importin-α is fundamental to eukaryotic cells as it shuttles cargo proteins into the nucleus through nuclear pore complexes (Görlich, 1997). However, more species-specific roles include proteins like the Radial spoke protein 14 (A8HNV0), needed for ciliary motility (Gui et al., 2021). In other species, the Apicomplexa phylum of parasites such as Plasmodium, the causative agent of malaria (Prevention, 2019),
have Armadillo-containing proteins predicted to have putative roles in invasion and motility, splicing, and environmental sensing that may be essential for its survival throughout the lifecycle (reviewed in Tewari et al., 2010).

Since Huber et al. (1997) solved the first structure of the β-Catenin protein (Figure 5.1C), many other structures of Armadillo-containing proteins have been shown to share a conserved three-dimensional structure. For example, canonical Armadillos in Figure 5.1A are 41 amino acids long with a conserved three α-helical structural motif: short helix with two turns (H1) and two longer helices H2 and H3 with 2-3 and 3-4 turns, respectively (Figure 5.1B). Helices H2 and H3 in Figure 5.1B lie antiparallel to each other (Huber et al., 1997), comparable to the helical packing of TPRs. In contrast, H1 folds perpendicular to the H2-H3 hairpin. The Armadillo repeats have a right-handed fold in Figure 5.1Ai (Kobe and Kajava, 2000). Figure 5.1C shows how adjacent Armadillos in an Armadillo-Domain, folds into a superhelix with a right-handed twist. Each successive Armadillo orientates 30° from the previous repeat, around an axis with little or no curvature (Kobe and Kajava, 2000, Groves and Barford, 1999).

Armadillo repeats are frequently compared to other helical repeat families such as HEAT (Andrade and Bork, 1995) and Pumilio () repeat families, due to their sequence and structural homology. HEAT repeats are probably evolutionarily related, as they share sequence and structural homology. However, HEAT repeats have multiple subfamilies that display structural variation. For example, kinks in the first helix of HEAT repeats resembles H1 and H2 in ARs (Andrade et al., 2001). In some cases, it is challenging to distinguish Armadillo from HEAT repeats because the structural variation amongst individual repeats of the same protein can belong to both families within the same Repeat-Domain. This is a specific feature unique to them, unlike the TPR superfamily that have only one type of TPR or TPR-like repeat in a Repeat-Domain. Furthermore, these repeats can sometimes be indistinguishable from one another, termed the twilight zone (Kippert and Gerloff, 2009).
In contrast, Pumilio repeats are evolutionarily conserved RNA-binding proteins. The topology of Pumilio repeats resemble Armadillo repeats, where Pumilio α-helix 1 resembles Armadillo H2, α-helix 2 with H3, and α-helix 3 with H1 (Wang et al., 2001). However, the Pumilio-Domains have an entirely different solenoid fold with a large curvature and no twist. In fact, the solenoid domain resembles to that of Ankyrin-Domains and Pumilio α-helix 2 makes up the concave surface.
Figure 5.1: Armadillo repeat sequence logo and structure. (A) Sequence logo of the Armadillo MSA in Figure 5.4. Amino acids coloured with ClustalX colour scheme. Generated with WebLogo on the online webserver tool. (B) The structural motif of an Armadillo repeat with three helices: H1, H2, H3. (Bi) Three helices depicted as cylinders with arrows highlight three helices and right-handedness of the polypeptide chain: H1, H2, H3. (Bii) Ribbon representation
5. Armadillo Repeat Family

5.3 Introduction

of 7-Armadillo-2-Domain (42-83aa) from the Designed Armadillo Repeat protein (PDB code: 4PLQ). The residues are coloured with the ClustalX colour scheme. (C) First solved structure (PDB code: 2BCT) of the Armadillo-containing protein called β-Catenin protein (Q02248) in Mus musculus (Huber et al., 1997). Alternating colours of teal and tomato for every adjacent repeat. The second panel of the same structure highlights helices H1, H2, and H3 with the colours green, blue, red, respectively. All turns are depicted as black lines. All structures here have been generated in UCSF Chimera.

The sequence logo summarising the Armadillo sequence in Figure 5.1B is characterised by a degenerate consensus sequence defined by a pattern of highly conserved positions with hydrophobic amino acids. Conserved positions 14, 17, 20, 21, 28, 32, 35, 36, 38, and 39 are occupied by amino acids such as leucine, isoleucine, valine, and alanine (Andrade et al., 2001). These positions, apart from 38, make up the hydrophobic core that runs the entire length of an Armadillo-Domain.

Armadillo-Domains form both a concave and convex surface. The ladder of H3 helices makes up the concave surface (Groves and Barford, 1999) in Figure 5.1C (orange cylinders). In contrast, the convex surface is the combined surface across H1 and H2 shown as green and blue coloured cylinders, respectively, in Figure 5.1C. Armadillo-Domains bind to substrates, by wrapping around them in an extended and antiparallel fashion to the succession of repeats, along the concave surface (Boersma and Pluckthun, 2011). Figure 5.1D displays some examples of how different Armadillo-Domains bind to various substrates.
Figure 5.2: Four examples of Armadillo-Domains and their interacting protein substrates. Listed from left to right: Junction plakoglobin (P14923, PLAK_HUMAN, Homo sapiens, teal) binding to Cadherin-1 (P09803, CADH1_MOUSE, Mus musculus, orange), PDB code 3IFQ;
5. Armadillo Repeat Family

5.3 Introduction

Catenin beta-1 (Q02248, CTNB1_MOUSE, Mus musculus, teal) binding to Adenomatous polyposis coli protein (P25054, APC_HUMAN, Homo sapiens, orange), PDB code 1JPP; Importin subunit alpha-1 (P52293, IMA1_MOUSE, Mus musculus, teal) binding to Polymerase basic protein 2 (P31345, PB2_I75A3, Influenza A virus, orange), PDB code 4UAF; Catenin delta-1 (O60716, CTND1_HUMAN, Homo sapiens, teal) binding to Cadherin-1 (P12830, CADH1_HUMAN, Homo sapiens, orange), PDB code 3L6Y. All Armadillo-containing proteins have been orientated from N- to C- terminus direction from the bottom up. All structures here have been generated in UCSF Chimera. The concave surface has a positively charged groove that is highly diverse to accommodate a wide range of substrates (Gul et al., 2017). Located at the bottom (fourth turn) of H3 on the concave surface the other conserved position, 38 in Figure 5.1B, forms an asparagine ladder. The asparagine residues bind the backbones of substrates, keeping them in an extended structure by forming bidentate hydrogen bonds (Conti et al., 1998). It is believed this conserved binding conformation is the primary role of asparagine ladders. The asparagine ladder (green) is demonstrated in Figure 5.2D in the Importin subunit alpha (teal) protein.

Importin subunit alpha in Figure 5.2D (coloured teal) belongs to a family of importin-alpha proteins, the members of which act as adapter proteins for nuclear pore complexes (NPCs). These proteins recognise the nuclear localisation signal (NLS) on cargo proteins to facilitate shuttling across the nuclear membrane (Lange et al., 2007). Additionally, there are conserved tryptophan residues on the third turn of H3 helices that form a ladder of residues within individual proteins. This tryptophan ladder is only observed in importin proteins that use both asparagine and tryptophan ladders to bind to the NLS (Conti et al., 1998, Conti and Kuriyan, 2000).

The ability of Armadillo-Domains to bind to extended polypeptide chains make them another promising repeat family to engineer into functional scaffolds that bind to a protein of interest with a range of affinities (Boersma and Pluckthun, 2011, Hansen et al., 2018). Parmeggiani et al. (2008) developed the first designed armadillo repeat protein (dArmRP) to demonstrate that stable Armadillo-Domain scaffolds can be designed from the consensus sequence. Madhurantakam et al. (2012) extended this work by employing consensus and
computational methods to develop a library of dArmRPs with varying curvatures that favour the modular substrate-binding reminiscent of natural Armadillo-Domains. Their study identified an optimal curvature suitable for continuous peptide binding that can be used as a template for developing modular peptide binders. (Reichen et al., 2016).

Armadillo-Domains were shown to organise themselves with a minimum of six adjacent Armadillos (Hatzfeld, 1998). Since that work, some reviews have collated the growing number of protein entries to understand the evolution and function of Armadillo-containing proteins. For example, Coates (2003) reviewed the different conserved subfamilies of Armadillo-containing proteins, and their functions, across different species. Furthermore, the arrangement of Armadillo repeats was visualised, showing that their PRA can be as short as a 4-Armadillo-Domain in the Phospholipase A I-like protein (F4HX14) or as long as a 42-Armadillo-Domain in the Protein Cellulose Synthase Interactive 2 (Q9C6Y4) protein of Arabidopsis thaliana (Coates, 2003). In contrast, Jernigan and Bordenstein (2015) measured the abundance of Armadillo-containing proteins across eukaryotes, bacteria, and archaea. Their study established that most bacterial and archaeal species have at least one Armadillo-containing protein, but they did not delve into the number of repeats. However, there has been no formal quantitative attempt to measure the PRA of Armadillo-Domains in the literature.

5.4 Methods
5.4.1 Gathering repeat annotations
On sampling Armadillo-containing proteins in InterPro, the InterPro Entry that best annotates on the Repeat level of Biological Entities is IPR000225. The contributing member databases of this InterPro Entry are SMART (SM00185), PROSITE (PS50176), and Pfam (PF00514). Explained in Section 2.3.1.2, viewing individual protein entries annotated by IPR000225 highlights how PROSITE has a different repeat annotation ruling to SMART and Pfam. Thus, only SMART and Pfam were selected for this analysis. In addition, all Armadillo annotations from Swiss-Prot called ‘ARM’ were also retrieved.
The annotations were compared to understand how comprehensively the different databases annotated Armadillo repeats. Figure 5.3 shows the relationship between all the annotations across SMART and Pfam protein signatures InterPro and the Swiss-Prot database. This plot highlights the discrepancy between annotations from different databases.

![Figure 5.3: Comparison of annotations from three databases. An UpSet plot that visualises how Pfam (PF00514), SMART (SM00185), and Swiss-Prot (“ARM”) databases annotate Armadillos in 325 Armadillo-containing proteins across 57 organisms.](image)

Finally, a non-redundant merge of all these annotations from the databases was carried out in the order: SMART, Pfam, and Swiss-Prot. This step generates 2267 annotations from 317 Armadillo-containing proteins across 57 species, and the dataset comprises 1587 SMART, 6 Pfam, and 674 Swiss-Prot annotations. Furthermore, SMART annotations were selected over Pfam because Pfam annotations tended to have truncated annotations of varying lengths, compared to SMART.

### 5.4.2 Multiple sequence alignment

The 2267 Armadillo sequences were aligned with the MULTALIGN program in AMPS. The template alignment for MULTALIGN is the Pfam family (PF00514) alignment of Armadillo sequences. Manual assessment of the template alignment removed 9 long sequences, leaving 188 sequences in the MSA. These non-canonical sequences were removed as they
made up columns that were insertions within helices and lacked structural data to support their inclusion. The helices are denoted above the template alignment in Figure 2.9A, where the secondary-structure-dependent gap penalties will be applied. The following parameters were assigned for the MULTALIGN program: standard gap penalty of 8 between helices; a gap factor of 100 represents a gap penalty of 800 within helices; a constant of 8; BLOSUM62 matrix. The quality of the MSA was visualised and assessed in the Jalview.

5.5 Results and discussion
5.5.1 Sequence analysis

As denoted above the MSA of Figure 5.4, RAM identified ten evolutionarily conserved positions ($PR_{Shenkin} < 25$). Studying the location of the most conserved positions in the structural motif, we observed how Helices 2 and 3 are more conserved than Helix 1 since all ten positions reside there. In Table 2.3, seven conserved positions display hydrophobic, aliphatic, and small properties. Supplementing this observation, Table 5.1 shows the top three most frequent amino acids for the five positions: leucine, isoleucine, and valine. For the other three positions, position 17 has leucine, isoleucine, and phenylalanine, and positions 20 and 21 have leucine, isoleucine, and methionine. Contrasting these seven conserved residues are two other positions with hydrophobic, small, and tiny properties. Position 32 has 54.54% occupancy with alanine, whereas, position 35 has 38.11% occupancy. Finally, the last conserved position, 38, is the characteristic site on the concave surface of Armadillo-Domains. This site is involved with protein–substrate binding and displays small, polar, and uncharged physicochemical properties, such as asparagine and serine.
5.5 Results and discussion

Figure 5.4: Armadillo MSA. The 28 sequences represent equidistant members of an average distance tree, calculated in Jalview from the full alignment. All annotations above and below the MSAs refer to information from the full-length MSAs. Annotation above MSA: The annotation ladder (black) numbers the alignment columns in intervals of 5 amino acids. Low occupancy alignment columns are hidden. Arrays of hidden columns are collapsed and depicted as a single blue line through the alignment with an arrowhead above the alignment. Annotation below MSA: four annotation tracks below alignments that summarise the information for the full-length MSA; PRShenkin, where lower scores indicate more conserved positions, and vice versa for larger values; secondary structure allocated based on summarised DSSP definitions across all Armadillo structures in Section 5.5.3.1; Glycine, refers to the absolute number of glycine amino acids at each alignment column; Proline, refers to the absolute number of proline amino acids at each alignment column.
The structure visualising the conserved positions in Figure 5.5 shows why the amino acid properties discussed above are observed at those nine hydrophobic sites. The aliphatic sites in Figure 5.5A and B occupy the gaps within and between Armadillos that make up the hydrophobic core running through the entire length of an Armadillo-Domain. The two positions dominated by alanine, in Figure 5.5C and D, reside in regions that do not permit larger hydrophobic sidechains. These conserved positions may indicate the importance of helices H1 and H2 for the structural fold of the Armadillo-Domain.

Table 5.1: Top three most frequent amino acids in each alignment column of the Armadillo MSA. The percentage values represent the proportion of the amino acid type present in an alignment column.
Alignment Column: High-occupancy alignment column in the MSA of Figure 2.9. Rank: The top three most frequency amino acids at each alignment position.

An additional four positions can be characterised for their helix breaking properties, contributing to the distinctive kink between H1 and H2. Below the MSA of Figure 5.4, additional bar plots show the proportion of glycine and proline separately. Glycine is the most dominant observed at positions 12 and 13, while proline resides at 14 and 15. The conservation for these amino acids can explain why H1 and H2 do not fold as a single helix, as seen in the evolutionarily related HEAT repeat family (Andrade et al., 2001).
Figure 5.5: Two adjacent Armadillos highlighting three amino acid classes of conserved positions. There are three hydrophobic amino acid classes: (A-B) aliphatic and small (blue), (C-D) small and tiny (red), and (C-D) polar and small (green). (E-F) Visualises all three classes together. Ribbon representation of the consensus-based dArmRP (PDB code: 4PQL; Reichen et al., 2014).
5.5 Results and discussion

5.5.2 Variant analysis

As discussed in Section 2.3.7.1 of the Methods, VarAlign mapped 5981 variants to human sequences in the alignment, and RAM identified positions constrained in the human population. Figure 2.18 displays the MES against PRShenkin for each position in the Armadillo MSA. Evolutionarily conserved CMD sites 17, 20, 21, 36, 38, and 39 are constrained in the human population. This step vindicates the classification for positions that may be key for the structural fold, visualised in Figure 5.6A. Position 38 is an exception, where it interacts with substrates on the concave surface (Conti et al., 1998, Conti and Kuriyan, 2000).

![Figure 5.6: Sequence and variant classifications.](image)

Contrasting CMD sites in Figure 2.18 are three CME positions. Counterintuitively, these positions are enriched for variants relative to the other conserved positions. The physicochemical properties of the variants at CMD site 20 and CME site 14 were studied to understand this observation. Figure 5.7 shows multiple substitution matrices showing amino
acids for the variants at selected alignment columns. For reference, Figure 5.7A shows how CMD site 20, conserved for leucine with 53.7% occupancy in Table 5.1, has 14 variants mutating to valine with similar physicochemical properties.

In contrast, CME position 14 displays different variant properties from position 20. Revisiting the MSA in Figure 5.4, Table 5.1 shows position 14 comprising leucine (30.5%) and isoleucine (25.1%) and valine (21.8%). The substitution matrix in Figure 5.7B shows that valine variants commonly mutate back to the leucine isoforms and vice versa. Other favourable amino acid substitutions include alanine and methionine. This substitution matrix highlights the presence of many conservative mutations with favourable physicochemical properties at that site. The CME classification may indicate that an Armadillo-Domain can permit different sidechains at CME sites without interfering with the structural fold. Whereas CMD sites within the hydrophobic core are limited, as seen with the constrained variant profiles.
5. Armadillo Repeat Family

5.5 Results and discussion

Figure 5.7: Substitution matrices for variants at four positions in the Armadillo MSA. Four positions represent classified sites from Figure 2.18: 20 (CMD), 14 (CME), 4 (UMD), 6 (UME). The y-axis represents the amino acid from the canonical sequence (original amino acid), and the x-axis represents the variant amino acid (new amino acid). The number of variants is numbered in each cell. Cells absent of variants are blank. Numbered cells filled in with a colour gradient from white (lowest value) to black (highest value) within each substitution matrix.
Unconserved positions in repeat families account for the bulk of the substrate interactions. Therefore, the variant analysis was used to elucidate what unconserved sites in Armadillo repeats may be constrained in the human population and are likely specialised for binding specificity.
Seven UMD sites with a $P_{Shenkin}$ of 75 or more and an MES scoring below the trendline were observed in Figure 2.18. The UMD sites were split into two categories, based on their location on the structure of Figure 5.6B. Positions 31, 34, and 41 were identified on the concave surface of H3 involved with substrate-binding. Therefore, this set of residues will be described as potential binding-specificity sites. In contrast, positions 4, 7, and 11 on the face H1 of a neighbouring Armadillo face away from the concave surface, and their low MES values indicate that they may have a structural role instead. Lastly, position 22 on H2 does not show a clear role but may have an important structural role, supported by the structural analysis.

Next, these sites were compared with UME sites that are evolutionarily unconserved, with little selective pressure, relative to other sites in the MSA. Apart from position 6 on H1, positions 16, 19, and 24 reside on H2. Since H3 is absent of UME sites, this stratification of unconserved sites likely stands true.

Like CMDs, unconserved sites were further explored for the amino acid properties of variants. Position 6, the most enriched for missense variants, was compared with position 4, which has the lowest MES from the unconserved sites. It is hard to discern the differences between the substitution matrices in Figure 5.7C and D since these sites are unconserved for many different amino acids. As a result, simply classifying these sites through MES alone is sufficient to identify the most important sites to investigate further in the structural analysis. Position 4 ($MES = -0.18, p = 0.09$) is most significantly depleted for missense variants amongst UMD sites that remains the most interesting in this family.

### 5.5.3 Structural analysis of evolutionary Conserved and Missense Depleted positions

ProIntVar retrieved 257 Armadillo-containing structures for 35 proteins, with 265 Armadillos across 10 species. RAM aggregated the structural features and generated the contact map for intra-Armadillo interactions in Figure 5.8A. The interactions between helices are focused on here to understand what residues are essential for the fold of an Armadillo repeat. First, there is an interaction between H2 (13–23) and the second linker peptides, with
H3 (28–39). A further interaction is H1 (5–10) with the tail end of H3 (38–39) as the Armadillo repeat folds back on itself.
Figure 5.8: Structural analysis of intra- and inter-Armadillo contacts. Each contact map is derived from 265 Armadillos from 35 Armadillo-containing proteins across 257 PDBe structures. (A) The intra- Armadillo contact map shows the residue interactions within each
Armadillo. (C) The inter- Armadillo contact map shows EoCs between adjacent Armadillo repeats. The x-axis to y-axis visualises contacts between alignment positions from the N- to C-terminus direction and vice versa. The alignment column axes represent equivalent positions in the Armadillo MSA of Figure 5.4. Cells are shaded according to the EoCs with the colour gradient bar to the right of each contact matrix. Both axes are paired with bar plots, showing the residue pair contacts observed in less than 50% of EoCs in intra- (89/178) and inter- (84/168) Armadillo residue contacts. The secondary structure of Armadillo repeats is depicted as boxes and linker regions as lines along both axes. The structure of two adjacent Armadillo repeats in dArmRP (Reichen et al., 2014), numbered 1 and 2, together with the secondary structure on the x and y-axes of the contact maps, orientate the residues involved with interactions (PDB code: 4PLQ). The individual positions that meet the average CN plotted on each bar plot along contact map axes have their sidechains visualised and coloured onto dArmRP for both (B) intra- and (D) inter- dArmRP contacts. (E) The combination of both interaction bar plots, with a threshold line of the average CN. (F) Median residue surface accessibility for each alignment position. The different bars represent the proportion of residues with structural coverage that fall into the different RSA classifications according to the specified thresholds: surface (RSA > 25) partially exposed (5 < RSA < 25) or core (RSA < 5).
RAM filtered out residue pair contacts observed in less than 89/178 EoCs (50%) and these were listed in Table 5.2. To understand the network of interactions involved with the intra-Armadillo structural fold, positions that meet the average CN (CN > 2) were visualised in the structure of Figure 5.8B. Positions 17, 20, 21, 36, and 39 from the seven CMD positions
have high CNs, whereas CMD site 35 did not. Next, CME site 32 also has a high CN. In contrast, position 29 (CN=4) was not inferred from earlier analysis.

Table 5.2: Residue pairs involved with intra-Armadillo contacts and their supporting EoCs. The contact pairs that are observed in the contact matrix of Figure 5.8A. All interactions listed here are residue pairs with more than 50% (89/178) EoCs.

<table>
<thead>
<tr>
<th>X-Axis</th>
<th>Y-Axis</th>
<th>EoCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>31</td>
<td>168</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
<td>168</td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>166</td>
</tr>
<tr>
<td>37</td>
<td>35</td>
<td>164</td>
</tr>
<tr>
<td>21</td>
<td>16</td>
<td>164</td>
</tr>
<tr>
<td>30</td>
<td>28</td>
<td>163</td>
</tr>
<tr>
<td>40</td>
<td>39</td>
<td>160</td>
</tr>
<tr>
<td>33</td>
<td>32</td>
<td>160</td>
</tr>
<tr>
<td>37</td>
<td>34</td>
<td>159</td>
</tr>
<tr>
<td>37</td>
<td>38</td>
<td>158</td>
</tr>
<tr>
<td>30</td>
<td>27</td>
<td>155</td>
</tr>
<tr>
<td>34</td>
<td>31</td>
<td>153</td>
</tr>
<tr>
<td>29</td>
<td>28</td>
<td>153</td>
</tr>
<tr>
<td>40</td>
<td>38</td>
<td>152</td>
</tr>
<tr>
<td>29</td>
<td>20</td>
<td>151</td>
</tr>
<tr>
<td>40</td>
<td>35</td>
<td>150</td>
</tr>
<tr>
<td>40</td>
<td>5</td>
<td>150</td>
</tr>
<tr>
<td>21</td>
<td>17</td>
<td>150</td>
</tr>
<tr>
<td>33</td>
<td>35</td>
<td>148</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>148</td>
</tr>
<tr>
<td>26</td>
<td>28</td>
<td>147</td>
</tr>
<tr>
<td>33</td>
<td>28</td>
<td>147</td>
</tr>
<tr>
<td>36</td>
<td>35</td>
<td>146</td>
</tr>
<tr>
<td>41</td>
<td>38</td>
<td>146</td>
</tr>
<tr>
<td>18</td>
<td>16</td>
<td>145</td>
</tr>
<tr>
<td>26</td>
<td>25</td>
<td>144</td>
</tr>
<tr>
<td>33</td>
<td>20</td>
<td>143</td>
</tr>
<tr>
<td>39</td>
<td>4</td>
<td>141</td>
</tr>
<tr>
<td>30</td>
<td>31</td>
<td>140</td>
</tr>
<tr>
<td>37</td>
<td>31</td>
<td>140</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>138</td>
</tr>
<tr>
<td>36</td>
<td>17</td>
<td>137</td>
</tr>
<tr>
<td>18</td>
<td>11</td>
<td>137</td>
</tr>
</tbody>
</table>
X-Axis: The alignment column on the x-axis of the contact map in Figure 5.8A. Y-Axis: The alignment column on the y-axis of the same contact map. Armadillos: The value representing the evidence of contact.

Next, an inter-Armadillo contact map was studied to understand the general SSE interactions between the helices: the slight twist of Armadillo-Domains permits the interaction between parallel helices H1$^1$-H1$^2$, H2$^1$-H2$^2$, and H3$^1$-H3$^2$ that are parallel to each other; helix H2$^1$ interacts with the second half of helix H1$^2$; The first half of helix H3$^1$ interacts with the complete length of helix H2$^2$, and the second half interacts with helix H1$^2$.

Next, sites with the highest CN (CN > 3) were identified in the inter-Armadillo contact map of Figure 5.8C, and those residue pairs are listed in Table 5.3. The data were initially split into forward and backward interactions to establish positions specialised in inter-Armadillo interactions. Forward interactions include positions 18, 21, 30, 33, 36, 37, 39, and 40, as

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>17</td>
<td>136</td>
</tr>
<tr>
<td>40</td>
<td>8</td>
<td>135</td>
</tr>
<tr>
<td>36</td>
<td>8</td>
<td>133</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>132</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>131</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
<td>127</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>126</td>
</tr>
<tr>
<td>21</td>
<td>13</td>
<td>122</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>118</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>110</td>
</tr>
<tr>
<td>39</td>
<td>1</td>
<td>106</td>
</tr>
<tr>
<td>36</td>
<td>13</td>
<td>104</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>103</td>
</tr>
<tr>
<td>39</td>
<td>8</td>
<td>103</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>98</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>98</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>94</td>
</tr>
<tr>
<td>39</td>
<td>5</td>
<td>91</td>
</tr>
<tr>
<td>41</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>30</td>
<td>25</td>
<td>88</td>
</tr>
</tbody>
</table>
highlighted in the structure of Figure 5.8D. In contrast, positions 4, 8, 28, 31, and 35 are involved with backward interactions (Figure 5.8D).

Table 5.3: Residue pairs involved with inter-Armadillo contacts and their supporting EoCs. The contact pairs that are observed in the contact matrix of Figure 5.8C. All interactions listed here are residue pairs with more than 50% (84/168) EoCs.

<table>
<thead>
<tr>
<th>X-Axis</th>
<th>Y-Axis</th>
<th>EoCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>13</td>
<td>168</td>
</tr>
<tr>
<td>33</td>
<td>31</td>
<td>168</td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>166</td>
</tr>
<tr>
<td>21</td>
<td>16</td>
<td>164</td>
</tr>
<tr>
<td>37</td>
<td>35</td>
<td>164</td>
</tr>
<tr>
<td>30</td>
<td>28</td>
<td>163</td>
</tr>
<tr>
<td>40</td>
<td>39</td>
<td>160</td>
</tr>
<tr>
<td>33</td>
<td>32</td>
<td>160</td>
</tr>
<tr>
<td>37</td>
<td>34</td>
<td>159</td>
</tr>
<tr>
<td>30</td>
<td>27</td>
<td>155</td>
</tr>
<tr>
<td>29</td>
<td>28</td>
<td>153</td>
</tr>
<tr>
<td>34</td>
<td>31</td>
<td>153</td>
</tr>
<tr>
<td>40</td>
<td>38</td>
<td>152</td>
</tr>
<tr>
<td>29</td>
<td>20</td>
<td>151</td>
</tr>
<tr>
<td>40</td>
<td>5</td>
<td>150</td>
</tr>
<tr>
<td>21</td>
<td>17</td>
<td>150</td>
</tr>
<tr>
<td>40</td>
<td>35</td>
<td>150</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>148</td>
</tr>
<tr>
<td>33</td>
<td>28</td>
<td>147</td>
</tr>
<tr>
<td>41</td>
<td>38</td>
<td>146</td>
</tr>
<tr>
<td>36</td>
<td>35</td>
<td>146</td>
</tr>
<tr>
<td>18</td>
<td>16</td>
<td>145</td>
</tr>
<tr>
<td>26</td>
<td>25</td>
<td>144</td>
</tr>
<tr>
<td>33</td>
<td>20</td>
<td>143</td>
</tr>
<tr>
<td>39</td>
<td>4</td>
<td>141</td>
</tr>
<tr>
<td>37</td>
<td>31</td>
<td>140</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>138</td>
</tr>
<tr>
<td>36</td>
<td>17</td>
<td>137</td>
</tr>
<tr>
<td>18</td>
<td>11</td>
<td>137</td>
</tr>
<tr>
<td>33</td>
<td>17</td>
<td>136</td>
</tr>
<tr>
<td>40</td>
<td>8</td>
<td>135</td>
</tr>
<tr>
<td>36</td>
<td>8</td>
<td>133</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>132</td>
</tr>
</tbody>
</table>
5. Armadillo Repeat Family

5.5 Results and discussion

X-Axis: The alignment column on the x-axis of the contact map in Figure 5.8C. Y-Axis: The alignment column on the y-axis of the same contact map. EoC: The value representing the evidence of contact.

The intra- and inter- contact data were combined in Figure 5.8E to identify positions with dual structural roles in the Armadillo-Domain. All CMD sites apart from positions 38 and CME site 14 and 32 had high CNs in Figure 5.8E. Moreover, positions 5, 8, 18, 29, 33, and 40 with high CNs were not highlighted earlier in the sequence and variant analysis. Therefore, these positions (CN > 4) can be split into three groups: the first group is 8 and 18, the second group is 33 and 40, the third group is 5 and 29.

In the first group, position 18 (MES = 0.12, p = 0.22) is marginally conserved and position 8 (MES = 0.01, p = 0.88) is not as conserved. Nevertheless, both sites share very similar variant profiles in the substitution matrices of Figure 5.7B with CME sites, like in the representative example position 14. Furthermore, these sites have varying proportions of the same amino acids in Table 5.1 to CME sites. These two sites were not selected previously due
to the stringent PR\textsubscript{Shenkin} thresholds, but the variant profiles and structural analyses support their importance in the Armadillo-Domain.

The second group, with positions 33 and 40, shares similar variant profiles (Figure 5.7H). Furthermore, these two positions display similar amino acids (Table 5.1) to positions 32 and 35 and thus belong within the same amino acid class of small and tiny (Figure 5.5C/D). Previously classified into the same conserved class in Figure 5.5C and D, positions 33 and 40 belong in the same category that has been structurally supported as important.

Lastly, the third group are residues important from the structural analysis but do not seem as important from the sequence and variant analysis. Although position 5 ($MES = 0.08$, $p = 0.42$) is less conserved than position 29 ($MES = 0.03$, $p = 0.77$), both sites have similar variant profiles in Figure 5.7E, due to the arginine they are conserved for (Table 5.1). Approximately $\frac{3}{4}$ of all arginine variants for both positions mutate to hydrophobic residues with larger sidechains, such as cysteine, tryptophan, and histidine, and the rest mutate to glutamine.

Although inferred to be important, CME sites 32 and UMD site 4 did not have high contact numbers. Position 32 is highly conserved for alanine, with a high contact number in intra-Armadillo interactions. Therefore, it is likely to be important only for the fold of the Armadillo repeat and not the Armadillo-Domain. The same can be said for position 4, with a high contact only inter-Armadillo contacts for backward interactions.

5.5.3.1 Median relative solvent accessibility

The RSA data was analysed across all Armadillo repeats to support the structural analysis further. Figure 5.8G visualises the median RSA at each position in the alignment. All positions inferred from the sequence and variant analysis or supported with structural analysis had median RSA values under 10.
5.5.4 Deciphering Armadillo interaction specificity with protein substrates

Positions not inferred as necessary from the structural analysis were revisited in the context of substrate interactions. RAM collated 170 Armadillos across 26 Armadillo-containing proteins that interact with 112 protein substrates and is summarised in Figure 5.9. The bar plot in Figure 5.9A shows the number of EoCs at each position in the alignment. The bulk of substrate interactions occurs at helix H3, which makes up the concave surface in Figure 5.9B.

Figure 5.9: Aggregated substrate interactions. All non-redundant substrate interactions with Armadillo repeats are represented as EoCs (See Section ). (A) Bar plot of EoCs. The black line at 37 EoCs represents the average EoC. (B) Ribbon representation of an Armadillo repeat from dArmRP (PDB code: 4PLQ). Residues are coloured by the number of EoCs using the colour
gradient to the left of the structures. Each residue has its sidechains displayed that directly face the concave surface.

UMD sites 31, 34, and 41 and CMD site 38 are enriched for substrate interactions (EoC > 100). However, UMD sites 4, 7, 11, and 22 are not missense depleted. There is little confidence in the classification of positions 7 and 22 as they are close to the residual line with little structural support for any real function in the Armadillo repeat family. Additionally, 11 (MES = -0.08, p = 0.45) has 3 CNs and is entirely on the surface in Figure 5.8G. However, the most depleted position 4 (MES = -0.18, p = 0.09) amongst unconserved sites is structurally supported with 4 CNs and marginally buried. This novel site may be important for the fold of the Armadillo-Domain.

In contrast, eight other sites are not identified from the sequence and variant analysis with many substrate interactions (EoC > 37). Positions 1, 2, 5, 27, 30, 37, 40, and 41 all have different sequence and variant profiles that may explain the observed substrate interactions.

Position 1 (MES = 0.03, p = 0.77) is neither conserved nor unconserved but is depleted for missense variants. Studying its amino acid properties, the top three amino acids in Table 5.1 are asparagine, serine, and aspartate, which are conserved for hydrophobic, small, and polar physicochemical properties. It is difficult to interpret the variant profile in Figure 5.7F for this site, but this site may have a minor role like position 38 as a binding affinity site.

Position 37, conserved for serine, arginine, and glycine, has a variant profile in which many arginines mutate to equal proportions of cysteine (11), histidine (11), and glutamate (10) in Figure 5.7G. The variants can be described as either polar or charged residues. However, the rest of the sites with many substrate interactions are either on the trendline or enriched for variants. The only role these sites may have is in binding affinity. Furthermore, these sites may have little-to-no role and may have many substrate interactions due to being so close to substrates.
5.6 Conclusions

5.6.1 Key insights

High-quality MSA with the AMPS alignment method improves the alignment of single repeats in the Armadillo repeat family. Explored in other repeat families, this method proves to be the best alignment method, due to the low sequence homology between repeats within the same family.

Variant analysis was complemented by characterising variants at each site. The substitution matrices visualised the variant profiles for sites with high contacts that were conserved and missense enriched (CME). This analysis highlighted the enrichment for other amino acids with favourable physicochemical properties. This analysis was extended to other positions to understand them further.

The structural analysis validated conserved sites depleted for missense variants (CMD) with high contact numbers, and the RSA data classify them as buried. Contrasting these CMD sites, position 38 was the only exception, which was enriched for substrate interaction. Whereas unconserved and missense depleted (UMD) sites 31, 34 and 41 were enriched in substrate interactions and may be accountable for binding-specificity. Finally, a novel UMD site 4 may have an important inter-repeat structural role in the Armadillo-Domain.
Chapter 6
Conclusions and future work

6.1 Summary

The work presented in this thesis focuses on studying human population variation data in repeat proteins with a comprehensive structural analysis of three repeat families in Class III. This was achieved by: 1) developing a novel and modular workflow to analyse repeat proteins; 2) to study the literature and gathering appropriate repeat annotations; 3) improving the current methods for building MSAs and quantitatively summarising MSAs; 4) introducing human population variation analysis to repeats; 5) quantitatively analysis all available structural data to increase the understanding of Repeat-Domain folds and substrate interaction. This chapter is an overview of the contributions made in each chapter and provides suggestions for future work.

6.2 Repeat Analysis Method

Chapter 2 describes the development of RAM and summarises the tools and methods implemented to study Class III repeat families. RAM provides a straightforward and consistent
methodology for studying repeats in proteins. The main features of RAM are: 1) gathering and comparing repeat annotations from different databases; 2) applying a quantitative PRA analysis; 3) implements existing alignment methods, together with one developed here; 4) integrating human population variation analysis to repeat MSAs, characterised variant profiles of alignment columns, and highlight positions depleted for missense variants; 5) aggregate and summarised all available structural data for sequences in the MSA; 6) provides an interactive environment with a series of Jupyter notebooks to gather annotations, alignment repeat sequences, study genetic variation, and structural analysis in protein repeat families.

RAM was designed to adopt new tools and methods due to its modular framework. With ever-growing collections of whole-genome sequencing data of humans, including other species, future efforts can focus on incorporating more variant data and develop new methods. Improvements in alignment methods can build better and more informative multiple sequence alignments. Developments in three-dimensional protein structure prediction methods can supplement existing structures by expanding the number of structures to study with and improve structural-based alignments. All these ongoing developments would continue to evolve this pipeline.

While this was designed for Class III repeat families, it can be easily extended to other repeat classes. The suggested nomenclature and methodology can be flexibly applied to these repeat families as well to gain insights into how all types of repeats evolve across the genome.

### 6.2.1 Gathering repeats

Section 2.3.1 discusses how to find the most appropriate InterPro Entry that best annotate the repeat family of interest. This includes the correct annotation ruling that represents the structural motif established in the literature. All member databases contributing to the InterPro Entry of choice, together with Swiss-Prot and any other databases, are compared to see the discrepancies in annotating the repeat family. Furthermore, this section also explains
how to assess the quality of the source database in terms of their coverage and how comprehensive their annotations are.

Obvious improvements at this stage of RAM would include other datasets outside of InterPro and Swiss-Prot. For example databases such as TPRpred (Karpenahalli et al., 2007) for TPR and TPR-like annotations, or REP2 (Kamel et al., 2021) to suggest annotations across eleven common Class III repeat families. With the sole aim of having the most comprehensive and high-confident set of repeat annotations from multiple sources, this would be a first task for future work.

An alternative to sequence-based detection methods are structure-based annotations. Protein structures remain the gold standard for studying protein domains. The RepeatsDB can be described as the Swiss-Prot of repeats. They provide structure-based repeat annotations but is not sufficiently large to use for individual repeat families.

There is currently a large archive of protein structures in the PDBe. With that said, improvements in protein structure prediction, such as AlphaFold2 or RoseTTAFold (Baek et al., 2021), have considerably expanded the collection of structural models available for repeat proteins. For example, the Database Scanning function with STAMP would be key to incorporate more sequences into the analysis. Section 1.4 describes in-depth how repeats that are divergent in sequence, maintain structural similarity. Acknowledging this, where methods fail to detect repeats on the sequence-level, this approach can fill in the gaps and integrate more diversity that sequence profiles alone struggle with. Here, maxHMM would be ideal to dissect the repeat family or superfamily, such as TPRs with TPR-like repeat families, and characterise them into their respective subfamilies.

6.2.1.1 maxHMM

MaxHMM was developed and implemented into RAM to help fill in some of the gaps between annotations from InterPro and Swiss-Prot. Section 2.3.6 discusses the multi-level HMM library implementation that was a first of its kind in the repeat field. This method
Conclusions and future work

6.2 Repeat Analysis Method

successfully filled in gaps with novel TPR annotations unobserved in both databases. With the TPR family, it better represented the individual TPRs in the conserved 3-TPR-Domains that bind to the MEEVD motif of Hsp70 and 90 (Scheufler et al., 2000). Generating subfamilies helped identify related families of TPR-containing proteins without a priori knowledge. Despite the successes however, the current version of maxHMM has several limitations that can be improved with suggestions discussed below.

6.2.1.2 Current limitations of maxHMM

The final section on maxHMM depends on the manual curation of matches and thus requires many steps that make it quite laborious. The advantage of a multi-level HMM library approach is that maxHMM can harness unique conservation profiles in subfamilies that make up diverse families.

Seed TPRs in Section 3.3.3.2 were considerably better represented, in terms of the bit score. Other matches scoring above the threshold were often TPRs that make up conserved 3-TPR-Domains. However, this strategy is not optimal for detect novel repeat sequences that do not belong to conserved evolutionarily related TPR-Domains. Since repeats considerably diverge with one another and evolve with their protein substrates, the library of HMM profiles struggle to detect novel repeats. Moreover, HMMer is unable to detect repeats with indels since they are rarely observed within the training alignment. Some of these issues poses a challenge for maxHMM and discovering repeats altogether. Consequently, below are a few suggestions for the improvement of maxHMM and help automate many of the steps in the current manual stage. These include ways to improve detecting of novel repeat sequences, while tackling indels in repeat sequences.

6.2.1.2.1 Diversifying the training dataset

To identify novel repeat sequences, some strategies have been experimented that have not been scaled up to entire repeat families. Nonetheless, they prove to have some success and may potentially prove critical for the future development of maxHMM.
6. Conclusions and future work

6.2 Repeat Analysis Method

The sequence of repeats are highly diverse, while preserved their structural fold. To understand how to incorporate sequence variation within HMM profiles is a logical step forward to making HMM profiles more sensitive to novel sequences. There are four ways to incorporate sequence variation here: 1) Training on repeat annotations from different species; 2) Building MSAs across the entire repeat superfamily, for example TPR and all TPR-like; 3) Generate artificial repeat sequences with amino acid substitutions from human population variation data or shuffling amino acids within each column of a subfamily MSA.

While Repeat-Domains are more conserved across species (Chen et al., 2017), there is still sequence variation that can be incorporated into the individual subfamily HMM profiles. Since maxHMM was only trained on human TPR sequences, it represents a small fraction in the evolution of species. Incorporate variation in sequence through other species greatly improves the variability in any one HMM profile, that one species alone couldn’t. The reason being, is that the same repeat in the same protein across multiple species has had different selective pressures, encapsulating the variability of that one repeat.

The first and simplest solution is to train on all species. Although the focus of the current work is on humans, the footprint represents a small fraction in the evolution of species. The best examples are the unicellular organisms like bacteria and archaea are ancient. Sampling random TPR-containing proteins in lower caste species have incredibly diverse proteins, let alone the behemoth proteins that extend beyond the average length of a protein littered with many TPRs. Incorporate variation in sequence through other species greatly improves the variability in any one HMM profile, that one species alone couldn’t. Future work would train on repeat sequences across multiple species.

In the same manner to training on multiple species, incorporating other repeat families in the superfamily may increase sequence variation. A few TPR-containing proteins also had TPR-like annotations, and vice versa. MaxHMM matches below the threshold identified
putative TPRs in proteins with TPR-like-Domains. Mittl and Schneider-Brachert (2007) superimposed an isolated TPR and SLR confirmed they are highly similar. The same goes for Armadillo and subfamilies of HEAT repeats (Andrade et al., 2001), where individual proteins have repeats folding from both repeat families (Kippert and Gerloff, 2009). Further lengthening the MSA with similar evolutionarily related repeat families may help introduce the diversity maxHMM needs to improve. Applying maxHMM to a repeat superfamily and testing this idea would be the way to go forward in the future.

A largely unexplored concept is to generate artificial sequences, or pseudo-sequences, to complement the training dataset. During the development of maxHMM, studying its performance in detecting novel sequences highlighted undetected repeats using solved structures to confirm this. A common issue was that the underlying sequence comprised of residues observed in the complete and subfamily MSA but were poorly detected by HMM profiles.

To understand why this was happening, when HMM profiles should theoretically detect them, undetected repeat sequences conserved in structure were clustered with the training MSA. The subfamily to which the undetected sequences belonged to were used as a source of generating pseudo-sequences. Undetected sequences comprised of residues observed across different sequences. Pseudo-sequences were generated by randomly selecting amino acids observed at each column within the subfamily MSAs with no weighting. The MSA of pseudo-sequences were applied through maxHMM and previously undetected sequences were discovered with high confidence.

This unorthodox approach demonstrated that the library of HMM-profiles trained on real sequences are not diverse enough detect novel sequences. However, a more intuitive and less erratic approach discussed, would be to take advantage of coevolving sites identified from the structural analysis (Section 2.3.8) and generate pseudo-sequences in this manner.
Another approach is to generate pseudo-sequences using the missense variants observed in human sequences. Repeat sequences with tolerable missense mutations from the human population would a sensible venture as a future direction. This is another strategy to introduce the diversity this method lacks.

These two approaches discussed here would help incorporate the sequence diversity maxHMM can train on. Thus, making the library of HMM profiles more akin to detecting putative repeats with a confident score. These suggestions would be an exciting direction to take and explore in repeats to develop this method further.

6.2.1.2.2 Increasing flexibility of HMM profiles

Indels within repeats are rare compared to the bulk of canonical repeats. However, maxHMM rarely detected TPRs with indels. To support HMM profiles identifying non-canonical repeats, one suggestion is to have introduce large arrays between SSEs for the entire alignment with the amino acid X. Experimented with example non-canonical TPRs, maxHMM managed to detect TPRs with large insertions, albeit a low bit score. HMM profiles managed to incorporate this information and thus make them more sensitive to non-canonical repeats. This is a very simple improvement for future development of this method.

This strategy assumes all indels occur between SSEs (Zvelebil et al., 1987), but sometimes there can be insertions where SSEs extend further. The current implementation of maxHMM does not accommodate for different lengths of repeats. A solution and major contribution to the methodology would be to incorporate information from predicted protein structure information to help matches extend to the correct length.

Secondary structure methods, like JPred (Drozdetskiy et al., 2015) would help identify where the SSEs start and end, in order to correctly extend maxHMM matches above the threshold. Additionally, matches can be modified further to help connect adjacent putative repeat matches and thus improve the contiguity of repeats within Repeat-Domains.
Recent developments from the 13th and 14th Critical Assessment of Techniques for Protein Structure Prediction (CASP) showed a considerable improvement in tertiary structure prediction. AlphaFold was placed first in the overall rankings of the 13th CASP in December 2018 (Jumper et al., 2021, AlQuraishi, 2019). Sampling many repeat-containing proteins from different repeat families, the improved AlphaFold version 2 confidently predicted arrays of repeats for Class III. While not quantitatively assessed, many putative TPRs discovered by maxHMM with no structures are observed in AlphaFold2 structure models.

Using various prediction methods for SSEs and tertiary structural folds should in theory drastically improve the sensitivity of HMM profiles. This would be an interesting area to develop in maxHMM, thus helping to further improve its ability to detect non-canonical repeats.

6.2.1.2.3 Summary
With the early successes of maxHMM in TPRs, it can be improved further by responsively and dynamically detecting repeats. This approach would dynamically accommodate for repeat sequences that do not conform to the canonical set of repeats available. A truly novel repeat-aware tool that can combine PRA, detecting individual helices using trained profiles, and supporting them with predicted protein structure information. Future directions would include focusing on developing and incorporating all the suggestions discussed above. One distinct area lacking in this thesis is for these new implementations to be further tested against other known methods for detecting repeats: TPRpred for TPRs, PPRs, and Sel1-like repeats; REP2 to study the performance across all eleven repeat families this method detects; the comparison of individual databases that detect repeats with single HMM profiles, such as SMART, PROSITE, Pfam, etc.

6.2.2 Repeat architecture
The key highlights of PRA analysis is the new and consistent language that can raise new questions, and lastly a novel tool to quantitatively assess the organisation of repeats.
The new language proposed here will help individuals who study single repeat proteins and whole repeat families. Additionally, this will benefit others understanding the relevant literature. Some of the new questions were demonstrated with TPRs. Creating three large subfamilies based on global repeat architecture (start, middle, end), unveiled new aspects of sequence analysis of repeat families. Next, the largest three subfamilies of TPRs comprised the conserved 3-TPR-Domains without a priori knowledge. PRA analysis is a useful diagnostic tool for assess the integrity of the dataset. For example, PRA analysis can raise selected proteins that have unusually short Repeat-Domains if they do not agree with the literature consensus.

Similar to some of the work done to understand the prevalence of repeat families across different species (Jernigan and Bordenstein, 2015, Kamel et al., 2021), future implementations of PRA analysis can supplement work studying the evolution and function of repeat proteins proteome-wide across kingdoms. This method would quantitatively capture what repeat proteins share similar PRA and understand if they have the same functions. Furthermore, the PRA would conveniently supplement PDA analysis and may become the norm for protein family analysis.

RAM currently implements PRA analysis that identifies the number of proteins with multiple Repeat-Domains. This analysis can be further adapted to study how many proteins implement more than one Repeat-Domain across different repeat families. For example, the TPR and ankyrin repeat-containing protein 1 (O15050). With that, a future implementation of PRA would be to understand the combination of multiple Repeat-Domains across different repeat families.

The PRA nomenclature could be further developed. Initially designed to avoid the clutter of too many numbers in a single reference ID, the position of the Repeat-Domain, relative other Repeat-Domains in a single protein can be incorporated. This may prove to be useful in text when a piece of work is characterising and referring to the PRA. To consolidate
the new repeat terminology and language remains full-proof, future work would entail individuals to test the PRA and assess how they use it in their work with repeat proteins.

Finally, RAM acting as a resource, the PRA component of RAM would prove useful for engineering scaffold repeat proteins. Either designing repeat proteins with PRA observed in the repeat family or exploring other PRAs that are not naturally observed.

6.2.3 Multiple sequence alignments

Both alignment methods employed by RAM have successfully improved the alignment of repeats with lengths. However, they are two schools of thought here. The recursive alignment algorithm in Section 2.3.3.1 expects indels to occur at any region of a repeat. That is why this alignment method does not assume this and recursively realigns sequences where it introduces gaps. Furthermore, the advantage of this approach is that it avoids over-alignment of repeat sequences, whereby abstaining from aligning unrelated regions of sequences.

The argument in favour of the second method in Section 2.3.3.2 is that many popular alignment methods are focused on accuracy (Gil and Fiser, 2018) rather than building informative MSAs. Due to limited conservation within repeat families, most standard methods fail with aligning repeats. Attempts to align repeats with standard and accurate methods supported this ongoing challenge in assessing the quality of an MSA. AMPS utilises secondary-structure-based gap penalties, correctly permits diverse amino acids within SSE, even at conserved sites. Often, it is structurally tolerable since repeats are structurally receptive to indels. While it may not be accurate with standard methods, it certainly is informative. But the one limitation of this alignment method is the assumption that the underlying sequences represent the repeat of interest, which can be resolved by tweaking the gap factor.

The proposed alignment approaches in Section 2.3.3 are most suited for aligning repeats that are adequate for downstream analyses. However, the increased availability of structure models, such as AlphaFold or RoseTTAFold mean sequence alignment methods can be buttressed. Structural alignments will prove to be invaluable here. Future work would entail
using the STAMP search method against a database of known repeat-containing protein structures. This database can initially be from the PDBe, before expanding towards predicted structure models. The structure-based sequence alignment can be manually inspected to assess the quality of the MSA and thus, understand if this can be confidently automated. Additionally, this step would be assessed for how this approach copes with aligning repeats that have indels since this strategy aims to deal with this as one of the issues encountered in the work covered here. Once amended, other repeats sequences can be aligned to the template structure-based MSA with AMPS, as described in Section 2.3.3.2.

6.2.4 Variation data

The availability of human population variant data is likely going to continue to increase, especially with drastically low sequencing costs. However, there are other species with growing variation data. An increase in sequencing hospital strains of bacteria (Salipante et al., 2015), mean bacteria can be studied on how they continue to evolve and thus, a useful resource for studying variation (Robinson et al., 2011) in bacterial repeat proteins. The same can be said for bovines with many different breeds, which brings about source of variation in another species (Weldenegodguad et al., 2019, Boitard et al., 2016). Future efforts mean other species will be included, and thus, incorporate variation data in the analysis of repeats and other proteins.

Implementing genetic variation would be a great resource for those designing scaffold repeat proteins. This data would complement consensus-designed approaches that tend to design scaffolds (Binz et al., 2005, Boersma and Pluckthun, 2011). Acknowledging alternative amino acids beyond an MSA of paralogous repeat sequences would permit novel combinations of amino acids that would help adjust the thermostability of scaffold proteins and adjust specific properties for its function. Since evolution has done this naturally, it would provide a useful resource for exploring this without artificially generating and screening functional gene variants.
6.2.5 Structural analysis

With more structural data, from solving structures with conventional methods and predicted proteins, there are more structures to incorporate into the structural analysis of RAM. Incorporating the data is necessary to integrate the data, but will in fact add little in the understanding of studying the contact data of repeats. Mentioned in Section 2.3.8, inferring the most frequently observed residue pair contacts from the entire dataset is no different from sampling a smaller proportion of the structural data. Thus, the expansion of data will add little to the current conclusions obtained from the analysis with solved structures.

In contrast, the depth of structural data is useful when studying substrate interactions. One of the bottlenecks in understanding repeat function are how they interact with substrates. This source of information continues to grow, but at a much slower pace than solving individual solenoid proteins. The field will depend on new collections of solved repeat-to-substrate structures by further understanding the different binding modes that all repeats can accommodate. Potentially group repeats with similar binding modes and understand whether they share a similar depletion for missense variants at sites important for substrate-binding.

The structural analysis for identifying residue pair contacts is a very crude approach. RAM considers a pair of residues are in contact if they only one inter-atomic interaction exists. Furthermore, RAM does not delineate the type of interactions (i.e. Van der Waals, ionic, etc.), nor does it identify what atomic interactions occur. This point is raised here since the asparagine ladder in Armadillo-Domains is important for backbone interactions with protein substrates. It would be useful to consider identifying this information to tease out the type of interactions that occurs on concave surfaces of Repeat-Domains covered in this thesis. Moreover, how do missense variants that substitute for serine amino acids at the asparagine ladder affect its binding properties.
6.2.6 Subfamilies from structures
Acknowledging the earlier work with Armadillo and HEAT repeats by Andrade et al. (2001), the authors clustered repeat structures to support their sequence-based classification scheme. There may be clearer classifications of the individual repeat families if distance-based methods were employed to cluster together repeats with similar folds. This may improve segregating related repeat families, for example TPR and TPR-like repeats, and Armadillo and HEAT repeats. This can be achieved by analysing all the structures in the PDBe, extended with predicted structures, like AlphaFold2.

6.3 Repeat families
Chapter 2 focused on developing RAM that would give a logical and consistent approach to studying repeat families. RAM was then applied to the TPR (Chapter 3), Ankyrin (Chapter 4), and Armadillo (Chapter 5) repeat families to further demonstrate RAM and apply new analysis that integrates the increased coverage of genetic and structural data. Each chapter raised new questions that RAM did not originally address at the time of their analysis. Elements of RAM were explored in detail for individual repeat families. Below are the key insights of RAM.

6.3.1 Sequence analysis
A quantitative sequence analysis helped understand the amino acid properties at each position. With studies like Mosavi et al. (2002) that did a consensus-based analysis of Ankyrin repeats, RAM complements standard amino acid scoring methods with quantitative summaries of physicochemical properties at each alignment column. Originally implemented to understand TPRs, which have very few conserved sites, this helped understand why the amino acids were observed with respect to the TPR-Domain. This helped categorise multiple positions into specific classes of amino acid types, which were supplemented with their position and sidechain clearance. Although it was not difficult to decipher this, this sequence analysis was also applied to the Armadillo repeat family to elucidate the cohort of positions making up the
hydrophobic core. Summarised physicochemical properties characterised these repeat families better than standard amino acid scoring methods.

6.3 Repeat families

6.3.2 Human population variant data

Exploring human population variant data in all three repeat families presented interesting insights into the evolution of repeat families. Additionally, the observations made for all three repeat families conformed with the global trends observed across all protein domains (Pfam) in the MacGowan et al. (2017) study.

The classification of CMD sites often represented conserved sites important for structure and function. Therefore, the stratifications with alignment sequence conservation and human population variation validated this method by confirming what is expected for conserved sites across species.

Other isolated observations were CME sites. While missense enriched sites would be considered unimportant, the sequence and structural analysis would contradict otherwise. Analysis of missense variants with substitution matrices determined how they were enriched for amino acid substitutions with favourable physicochemical properties. In fact, these positions coincided with amino acid properties frequently observed across all species in the alignment. Other observations were conserved glycine or proline-induced turns between SSEs. These positions were also enriched for charge and hydrophilic amino acids. These observations may stand true for all CME sites across all Pfam families. Thus, it would be useful to extend this analysis to the rest of the repeat families, apart from the Armadillo repeat family in which this was extensively analysed.

Identifying potential substrate-specificity binding sites by dissecting positions with human constraint proved to be an exciting novelty in the repeat community. This approach further subset positions from similar sets of residues that Magliery and Regan (2005) have investigated with TPRs and Ankyrin repeats, which were the same positions identified from the structural analysis in RAM. All UMD sites were identified on the concave surface. UMD
sites were confirmed to be enriched for substrate interactions with the structural analysis. The only exception was position 4 in Armadillo repeats, which does not interact with substrates. However, studying the substitution matrices alone were not sophisticated enough to explain the variant properties at inferred binding-specificity sites. This may not be the way to support them. Therefore an alternative method should be considered when identifying them from UMEs. This point was made to acknowledge that UMEs are involved with substrate interaction, proving difficult justify them less critical from UMDs.

Although these classifications are vindicated in the structural analysis by aggregating structural data, this invites the opportunity for validating them. Briefly discussed in Section 3.4.2.2, the 3-TPR-Domains are the most conserved TPR-Domain in its family. The easiest way to validate UMD classifications computationally is to apply the same method with a different alignment. After aligning all 3-TPR-Domains, in theory, the additionally conserved positions 5, 6, 9, and 13, should remain missense depleted. Thus, these select UMD sites will transform to CMDs in this context.

To experimentally validate such positions, this would be limited to individual proteins. Using the same example, select variants at UMD sites would be mutated and the binding affinity of the MEEVD of 3-TPR-Domains tested and compared with wild-type. Using a similar selection process as in Llabrés et al. (2020), only conservative variants with a high allele count would be selected initially. These would act as critical control mutations to understand how the selective pressure acts on these sites. Additionally, they would provide the background mutation that occurs in the human population within this subfamily of TPR-Domains.

Rare (low allele count) variants with nonconservative mutations and pathogenic variants would be the next set of candidate mutations at UMD sites. Pathogenic variants are expected to display changes in binding affinity to the motif, whereas rare variants may be novel
mutations that may be pathogenic if they display similar properties in binding affinity to pathogenic variants.

The originality of this approach is using mutations tolerated in the human population rather than an alanine scan, for example, which gives very little information. This approach is much more sensible and rational than mutating positions randomly. Although this experiment outline has not been applied before, the study by Llabrés et al. (2020) uses carefully selected missense variants from the population as controls when interpreting the effect of pathogenic variants on the nanospring constant of the OGT TPR-Domain.

The same could be done for CMD sites, however it will not be as straightforward. Since mutations do not necessarily result in a misfolded Repeat-Domain, select mutations should be inserted and expressed in a cell line to study their effects with respect to their cellular function.

6.3.3 Structural analysis
A comprehensive structural analysis of this calibre has never been achieved in either of these repeat families. The ability to bring all available structures and dissect the roles of each position introduced a huge breadth of information about the structural fold of Repeat-Domains. In particular, the work here captures positions involved with intra- and inter-repeat contacts. All structurally and functionally important sites agreed with sites constrained in the human population.

Despite the breadth of structural data, bringing together so many structures proved to be less informative than the volume of data. Since repeats are modular, any more structural data adds little to the final conclusions. Although this point was not demonstrated, a small sample of TPRs with solved structures produced similar conclusions to the entire structural data, highlighting the highly modular properties of TPRs and other repeat families.

Unconserved sites depleted for missense variants (UMD) identified potential substrate-specificity binding sites. This was achieved by dissecting sites under human constraint, compared to other unconserved sites. The breadth of structural data proved to be insightful in
the understanding of binding modes that Repeat-Domains employ to interact with protein substrates.

Although it is informative to aggregate PPI data as RAM does, it is important to note that not all repeats are employed to bind to protein substrates. This may differ based on how the Repeat-Domain and protein substrate have evolved. Additionally, different substrates may have different binding modes to the same Repeat-Domain. Therefore, it would be informative to group repeats that share similar binding mode. This analysis may help tease out some specificity-binding sites and differentiate them further from other unconserved sites that are under little evolutionary constraint.

6.3.4 Pathogenic variants
Globular domains can fold and function as single entities. In contrast, repeats cannot fold by themselves without the support of adjacent repeats. While pathogenic variants were enriched at missense depleted sites (MacGowan et al., 2017), this analysis has proven to be less fruitful in repeats. Therefore, studying pathogenic variants in MSAs that align repeats in a tabular fashion may be the wrong format for understanding their effects.

Despite the sequence and structural similarity, mutations at the same position in different repeats within the same protein are not equivalent (Anna et al., 2021). Some mutations in TPRs can cause complete loss of function (Flom et al., 2006), but single or multiple mutations at highly conserved positions may have little to no effect. This is evident from the sequence variation tolerable across conserved sites across species and the missense variants observed in the human population. Counterintuitively, mutations at highly divergent sites can in fact result in detrimental effects. More specifically, the loss of localised binding interactions (Zeytuni and Zarivach, 2012, Voronin and Kiseleva, 2008, Perez-Riba and Itzhaki, 2019, Magliery and Regan, 2005) rather than unfolding of the protein (Kobe and Kajava, 2000, Llabrés et al., 2020). No single position is crucial, where in fact adding or removing entire repeats may not alter its function. An alternative approach is to study them in the context of
the whole Repeat-Domain (Laskowski et al., 2016, Llabrés et al., 2020). Finally, another limitation of using large databases like ClinVar and HMGD is that they do not archive new mutations that are only observed in published work within research communities studying specific repeat proteins. Therefore, this analysis is further missing pathogenic variants to understand them in the context of an MSA.

The results here with respect to sequence analysis and genetic variation affirm that repeats prove to be a very robust structural fold. Its ability to tolerate mutations are context dependent. Unfortunately, interpreting the effect of pathogenic variants is limited to experimentally characterising them or studying them through computational modelling. While building MSAs of repeat sequences has drastically increased the number of population variant data, this has instead made it difficult to interpret pathogenic variants.
## Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A-TPR-B-Domain</strong></td>
<td>Term to refer to a specific TPR in a TPR-Domain, where B denotes the numerical position of that TPR. ‘TPR’ can be any other repeat family, such as Ankyrin and Armadillo.</td>
</tr>
<tr>
<td><strong>A-TPR-Domain</strong></td>
<td>Two or more adjacent TPRs in a TPR-Domain, where A denotes the number of TPRs that comprise it. ‘TPR’ can be any other repeat family, such as Ankyrin and Armadillo.</td>
</tr>
<tr>
<td><strong>Alignment Column Occupancy</strong></td>
<td>Residue occupancy in a column of an alignment. An alignment column can be described as a high- and low-occupancy columns where residue occupancy is above or below 50%, respectively.</td>
</tr>
<tr>
<td><strong>Annotation coverage</strong></td>
<td>The number of proteins annotated by a given InterPro Entry.</td>
</tr>
<tr>
<td><strong>Annotation ruling</strong></td>
<td>How an annotation is defined with respect to the order of SSEs and its underlying sequence.</td>
</tr>
<tr>
<td><strong>Backward interactions</strong></td>
<td>Residue pair interactions between two repeats in the C- to N-terminus direction.</td>
</tr>
<tr>
<td><strong>Biological entity</strong></td>
<td>Term given to describe how InterPro Entries are classified into different types by InterPro curators. Homologous superfamilies, Families, Domains, Repeats and Sites are the five Biological Entities.</td>
</tr>
<tr>
<td><strong>Canonical repeat</strong></td>
<td>The most frequently observed length of a repeat in a repeat family with a clear structural motif.</td>
</tr>
<tr>
<td><strong>Capping repeat</strong></td>
<td>Repeats on the N- or C- terminus end of a Repeat-Domain.</td>
</tr>
<tr>
<td><strong>Comprehensive Annotations</strong></td>
<td>Following a proteome-wide PRA analysis, this term is used to describe a well annotated array of repeats that lacks gaps between repeats that are comparable to a canonical repeat.</td>
</tr>
<tr>
<td><strong>Domain</strong></td>
<td>Region of a polypeptide chain that can independently fold into a stable three-dimensional structure with a functional role.</td>
</tr>
<tr>
<td><strong>Domain Entity</strong></td>
<td>One of five hierarchical levels that make up Biological Entities. Domain Entities label regions of proteins that fold into a protein domain (See Domain).</td>
</tr>
<tr>
<td><strong>Extended hit</strong></td>
<td>Edited hits that have been extended to the length of the HMM profile it was detected with.</td>
</tr>
<tr>
<td><strong>Family Entity</strong></td>
<td>One of five hierarchical levels that make up Biological Entities. Family Entities describe and annotate whole proteins that share a common evolutionary origin.</td>
</tr>
<tr>
<td><strong>Forward interactions</strong></td>
<td>Residue pair interactions between two repeats in the N- to C-terminus direction.</td>
</tr>
<tr>
<td><strong>Hidden Markov Model</strong></td>
<td>Statistical representation used to model one-dimensional discrete symbol data. HMMs can capture the Markov states of the features in an MSA to model protein sequences, termed HMM profiles.</td>
</tr>
<tr>
<td><strong>Hit</strong></td>
<td>HMMer suite, which refers to a region in the sequence that the hmmbuild function detects with an HMM profile.</td>
</tr>
<tr>
<td><strong>HMM Profile</strong></td>
<td>Statistical representation of an MSA by summarising amino acid type feature at each alignment column in a Markov state.</td>
</tr>
<tr>
<td><strong>Homologous Superfamily Entity</strong></td>
<td>One of five hierarchical levels that make up Biological Entities. Homologous Superfamily Entities label proteins with a common evolutionary origin, as indicated by their structural similarities.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Indel</td>
<td>Insertion or deletion refers to the gain or loss of one or more bases within a coding region of the genome.</td>
</tr>
<tr>
<td>InterPro Entry</td>
<td>One or more member database signatures that represent the same Biological Entity.</td>
</tr>
<tr>
<td>Match</td>
<td>An extended hit that overlaps with one or more annotation that meets the Jaccard Index threshold.</td>
</tr>
<tr>
<td>Missense variants</td>
<td>Single base pair substitution that alters codon and translates to a different amino acid.</td>
</tr>
<tr>
<td>Pathogenic variants</td>
<td>A variant that increases the individual’s susceptibility or predisposition to a disease or disorder.</td>
</tr>
<tr>
<td>Protein Signature</td>
<td>The predictive models that represent tools to predict the presence of functional regions in protein sequences.</td>
</tr>
<tr>
<td>Raw hit</td>
<td>Hits that have not been processed or edited.</td>
</tr>
<tr>
<td>Repeat</td>
<td>Identical stretches of adjacent sequences that fold into a distinct structural motif and are organised in arrays.</td>
</tr>
<tr>
<td>Repeat-Domain</td>
<td>General term to refer to adjacent repeats that fold into a domain.</td>
</tr>
<tr>
<td>Repeats Entity</td>
<td>One of five hierarchical levels that make up Biological Entities. Repeat Entities label regions with repeats (see Repeat).</td>
</tr>
<tr>
<td>Scaffold proteins</td>
<td>Can interact with and bind to multiple members of a multi-complex protein in a signalling pathway that regulate signal transduction.</td>
</tr>
<tr>
<td>Sites</td>
<td>One of five hierarchical levels that make up Biological Entities. Site Entities label very short regions (often few amino acids), such as active sites, binding sites, post-translational modifications.</td>
</tr>
<tr>
<td>Structural Coverage</td>
<td>The number of residue pair interactions observed across a collection of structures.</td>
</tr>
<tr>
<td>Structural motif</td>
<td>The arrangement and combinations of SSEs making up a canonical repeat in structure.</td>
</tr>
<tr>
<td>Synonymous variants</td>
<td>Single base pair substitution that alters codon but translates the original amino acid before the mutation.</td>
</tr>
<tr>
<td>TPR-Domain</td>
<td>Two or more TPRs that meet the ( S \leq T ) condition, where ( S ) and ( T ) denote spacing and threshold, respectively, that are measured in amino acids. ‘TPR’ can be any other repeat family, such as Ankyrin and Armadillo.</td>
</tr>
<tr>
<td>UpSet plot</td>
<td>Visualising intersecting sets that is alternative way to Venn diagrams. The sets compared are visualised as a horizontal bar plot, the intersection sets are visualised as a matrix layout, and the number of elements in an intersection are represented as a vertical bar plot.</td>
</tr>
</tbody>
</table>
Bibliography


BARTON, G. 2002. OC-A cluster analysis program. *University of Dundee, Scotland, UK.*


PREVENTION, C. 2019. for DC and. CDC-Malaria-About Malaria-Biology.


receptor suggests roles for intrinsic flexibility and rigid domain motions in ligand binding. *BMC structural biology*, 7, 1-12.


