This thesis is submitted for the degree of
Doctor of Philosophy

Computational discovery and modelling of tandem domain repeats in proteins

Aleix Lafita Masip

European Bioinformatics Institute
University of Cambridge

Darwin College
February 2021

Declaration

I hereby declare that this thesis entitled "Computational discovery and modelling of tandem domain repeats in proteins" is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution.

I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution.

This thesis does not exceed the prescribed word limit of 60,000 words defined by the Degree Committee for the Faculty of Biology.

Aleix Lafita Masip
February 2021

Abstract

Computational discovery and modelling of tandem domain repeats in proteins

Aleix Lafita Masip
February 2021, Cambridge

Domains are functional and evolutionary units of proteins that typically fold into stable globular structures. A small subset of natural multidomain proteins contain large arrays of nearly identical domains repeated in tandem, challenging some of our assumptions about protein folding and evolution. In this study, I aim to discover new tandem domain repeats, characterise their sequence and structural properties and understand their roles in the function of proteins.

I start by using computational sequence analysis tools across large datasets of proteins and bacterial genomes to survey the prevalence and distribution of tandem domain repeats across organisms and domain families. Next, I computationally analyse and compare structures of domains found as tandem repeats, several of which have been experimentally determined by our collaborators in the course of this study. I finally develop two computational methods to systematically model the structure and misfolding energetics of tandem domain repeats.

Nearly identical tandem domain repeats are rare in natural proteins (below 0.1%) and their sequences are highly biased in amino acid composition. Many of them have structural roles in bacterial surface proteins implicated in biofilm formation and host colonisation; new examples of such proteins, named "Periscope proteins", show rapid domain repeat number variation, a molecular mechanism used to modulate bacterial phenotype. Tandem domain repeat structures reveal unusual structural malleability, with numerous cases of domain atrophy (loss of core secondary structures) and elaboration. They are also predicted to be more resistant to misfolding via tandem domain swapping, with potential misfolding-resistant mechanisms such as the domain topology and length.

This study improves our understanding of the prevalence, type and function of tandem domain repeats in proteins, in particular their role as structural elements in bacterial surface proteins, and suggests new protein and domain targets for further experimental characterisation. It also has important implications for protein misfolding and for the design and engineering of multidomain proteins.

Acknowledgements

This thesis would not have been possible without the support, advice and contributions of many other people that I have been fortunate to meet over the past four years.

I could not have wished for a better PhD supervisor than Alex Bateman. Through four intense years and over 130 one-to-one weekly meetings (I have counted 135), we have openly discussed about science, stared at large multiple sequence alignments to build new protein families, and assembled protein toy models by hand. I learned to pay attention to details, but also to see the big picture and to be pragmatic at times. I will always be grateful to Alex for caring so much about my work and responding to my questions so promptly, and for helping me grow as a scientist.

I have been extremely lucky to work alongside wonderful colleagues in our small research group. I thank every student, intern and visitor who has been part of the group in the past four years, specially Ananth and Matt for guiding me early in the PhD and Vivian for being a fantastic coworker and scientific ally.

The research lab of Professor Jennifer Potts have been terrific collaborators and instrumental to this work. I thank Jen for acting as a second PhD mentor and giving me support and advice throughout the project; and Michael, Fiona, Rachael and Sam for their important contributions. I am also grateful to our collaborators Robert Best and Pengfei Tian for catalysing and contributing to our work on tandem domain swapping, and Matthew Bowler for letting us choose our favourite protein domains to solve experimentally. I thank my EMBL-EBI colleagues Alex Almeida and Martin Hunt for their help and advice with Illumina and PacBio sequencing data; Tanmay Bharat and his group for useful discussions on the CdrA protein; and Andrey Kajava and his group for help with T-REKS.

I am grateful to the members of my Thesis Advisory Committee — Janet Thornton, Laura Itzhaki and Nassos Typas — for dedicating their valuable time to

discuss the project with me, always giving me positive feedback and constructive criticism. I thank EBI group leaders Nick Goldman, Pedro Beltrao and Zamin Iqbal for showing interest in my project and discussing ideas with me. I further thank anonymous journal reviewers for volunteering to read and provide valuable criticism of my work.

The EMBL Predoc community has been one of my best PhD experiences. It has been a great source of inspiration and support to get to know so many young and talented scientists; I thank everyone who made EMBL so special. I thank Raj and Vasileios for being fabulous friends and hosts, and fellow EBI Predocs from my batch — Ally, Borgthor, Conor, Jose and Michael — for making the EBI the coolest outstation. I also thank Ricard for being my Catalan-Basketball-EBI buddy in Cambridge and for his PhD advice.

I have been very fortunate to work at the EMBL-EBI and the Wellcome Genome Campus, it is a truly unique environment full of amazing scientists. I thank everyone in the Protein families team for our daily group lunches, the socials, the volleyball games, and their help with database questions, specially: Sara, Gustavo, Gift, Lorna, Ioanna, Typhaine, Blake, and Boris. I also thank all the Catalans on campus for making me feel at home: Aurelien, Eloy, Jordi, Alba, and specially Ruben and Laura. I thank Amy, the EMBL-EBI Research Office, the EMBL Graduate Office and the EMBL Career advisors for their support, making admin work easy, and for organising courses and activities for us.

I thank Darwin college and the DCSA for their support, and players of the Darwin Basketball club, specially Dan and Chris. I further thank the Cambridge University Basketball club and the Blues team for all the fun we had at practices, memorable games and celebrations.

I thank my former PSI colleagues and friends, Spencer and Jose, for our sporadic virtual meetings and for finishing together important projects started with Guido. I also thank Peter Rose for inviting me to come to San Diego and visit the UCSD.

This thesis is dedicated to my family — Guadalupe, Xavier and Carme — for their unfailing support and for always being there, and to my beloved partner, Annabel, who has always been by my side whatever the circumstances.

to Xavier, Guadalupe, Carme and Annabel

Contents

List of Figures

List of Tables

Preface

This work is the result of four years of research at the European Bioinformatics Institute (EMBL-EBI) in Cambridge, UK. I met Alex Bateman, my supervisor, at the EMBL-EBI PhD recruitment interviews in Januray 2017. He proposed me to work on a new research project in his group to investigate some unusual properties of protein domains that form long arrays of tandem repeats, and to collaborate with the structural biology group of Prof. Jennifer Potts at the University of York, who had recently solved experimental structures of four of these domains in bacterial surface proteins. I thought the project sounded very interesting and had potential for new and unexpected results, and it fitted with my scientific interests in protein structure modelling and evolution.

Some months later, in May 2017, I moved to Cambridge to start my PhD at the EMBL-EBI. The first task at hand was to find how prevalent these tandem domain repeats are in natural proteins, and to compile various protein sequence and structure datasets for future analyses. At the end of May, I presented some of these preliminary results to our experimental collaborators in an internal meeting at the University of York. The lab of Jennifer Potts had solved the structure of various domain repeats in *Staphylococcus* and *Streptococcus* surface proteins implicated in infection and biofilm formation. At the meting, I understood the roles of these repeats in bacterial proteins, and I was excited to see that some of their structures were atypical globular folds, which at first sight did not seem to be related to any other known protein fold. Upon closer inspection, Alex and I realised that these domains were in fact related to other widespread protein domains, namely Immunoglobulin and Ubiquitin folds, but had undergone large structural deletions of core secondary structures, a rare evolutionary event known as domain atrophy originally studied by Ananth Prakash, a former PhD student in our group. One of the domains, named Rib, was of particular interest because homologous sequences in its protein family contained large insertions at the presumed atrophied

region and therefore seemed to retain the complete Immunoglobulin structure. The lab of Jennifer Potts confirmed experimentally that these longer Rib domains fold into complete Immunoglobulin structures, revealing the most evolutionary recent domain atrophy case known to date. We published this story in the *PNAS* journal (I was a co-first author), and results are presented in Chapter 4 of this thesis.

In September 2017, I attended a conference to commemorate Prof. Jane Clarke's retirement at St. Catherine's college named PFEI (Protein Folding, Evolution & Interactions) featuring eminent scientists in the field from Cambridge and abroad. At the conference I met Dr. Robert Best, a researcher at the National Institutes of Health (NIH) in the USA, who had recently published a computational study on misfolding in tandem domain repeats using molecular simulations together with his student Dr. Pengfei Tian. I had previously contacted them to try to use their approach on other proteins, and we decided to start a collaboration to develop a computational method that would allow us to study more systematically tandem domain swapping, a type of protein misfolded conformation, across large datasets of protein domain structures. We released our method as an open source tool, published in a short *Bioinformatics* article, and reviewed the field and our perspective on tandem domain swapping in *Current Opinion in Structural Biology*. Outcomes of this work are described in Chapter 5.

Close interactions with Prof. Jennifer Potts and her research group at the University of York continued over the following years (2018-2019). Through our search for homologous bacterial surface proteins, we realised that the numbers of tandem domain repeats were extremely variable across genomes and that this variability could be used by bacteria to modulate their surface appearance and interactions. Jen proposed to name these proteins as "Periscope proteins" for their mechanistic resemblance to the submarine instrument. I set out to discover new Periscope proteins and systematically characterise their domain repeat number variability in bacterial genomes, but I quickly realised that the high sequence identities of domain repeats was problematic for assembling these Periscope genes by short-read sequencing technologies such as Illumina. I came across the NCTC3000 project, a collection of several thousand bacterial strains sequenced with PacBio long-read technology, which allowed us to reliably count the numbers of repeats in Periscope proteins and carry out a comprehensive analy-

sis of their variability. This study developed for over a year and I posted a preprint on *bioRxiv* at the end of 2020, which has been recently accepted to be published in the *PNAS* journal. Results are described in Chapter 3.

Over the course of the project, Alex and I had been compiling a list of interesting tandem domain repeats from our large-scale database searches and Pfam classification efforts, some of which had unknown structures. In January 2019, we joined a meeting of the EMBL Infections Diseases working group in Hamburg, where I met Dr. Matthew Bowler from EMBL Grenoble. Matthew was thinking about prototyping a sequence to structure determination service using the facilities developed at EMBL Grenoble, and we discussed the possibility to start a pilot project to solve some of the structures of bacterial surface domains from our list. I selected a subset of ten target domains and sent them to Matthew, who ordered the constructs, produced and purified the proteins using EMBL facilities and started crystallisation trials. Unfortunately, the project had to be suspended in March 2020 due to the COVID-19 pandemic, at the worst possible moment, when we had obtained the first protein crystals and were ready to start data collection at the ESRF Synchrotron. We hope to be able to resume it soon.

Several problems encountered during the course of this project required modelling protein structures in non-traditional ways, for example large structural deletions and misfolded conformations of proteins. Towards the end of the project, at the beginning of 2020, I came across a series of old articles from the 1960s describing a concept called "Euclidean distance matrices" (EDMs), which can be used to represent and manipulate points in space such as protein atomic coordinates. I realised that, by using protein distance matrices, these modelling problems could be solved by simple matrix rearrangements, and I implemented a fast and lightweight modelling approach for these types of structural rearrangements using EDMs. This work was published in the *F1000 Research* journal, and outcomes are described in Chapter 6.

I have presented results of my work in local and international meetings, including a keynote lecture at the EMBL Lab Day 2018 and poster presentations at the Protein Society conference in Seattle (USA, 2019) and the ISMB-ECCB in Basel (Switzerland, 2019).

Aleix Lafita Masip
February 2021, Cambridge

Published work

Parts of this work have been published or submitted to scientific journals, as indicated in the text for each chapter (* stands for equal contribution):

Chapter 2

V Monzon, **A Lafita** & A Bateman. Discovery of Fibrillar Adhesins across Bacterial Species. Accepted in *BMC Genomics* (2021). *bioRxiv*: https://doi.org/10.1101/2020.12.07.414375

Chapters 3 & 4

F Whelan*, **A Lafita***, SC Griffiths*, REM Cooper, JL Whittingham, JP Turkenburg, IW Manfield, AN St. John, E Paci, A Bateman & JR Potts. Defining the remarkable structural malleability of a bacterial surface protein Rib domain implicated in infection. *PNAS*. 116 (52), 26540-26548 (2019). https://doi.org/10.1073/pnas.1911776116

F Whelan*, **A Lafita***, J Gilburt*, C Degut*, SC Griffiths*, HT Jenkins, AN St John, E Paci, JWB Moir, MJ Plevin, CG Baumann, A Bateman & JR Potts. Periscope Proteins are variable length regulators of bacterial cell surface interactions. Accepted in *PNAS* (2021). *bioRxiv*: https://doi.org/10.1101/2020.12.24.424174

Chapter 5

A Lafita, P Tian, RB Best & A Bateman. TADOSS: computational estimation of tandem domain swap stability. *Bioinformatics*. 35 (14), 2507-2508 (2019). https://doi.org/10.1093/bioinformatics/bty974

A Lafita, P Tian, RB Best & A Bateman. Tandem domain swapping: determinants of multidomain protein misfolding. *Curr Opin Struct Biol*. 58, 97-104 (2019). https://doi.org/10.1016/j.sbi.2019.05.012

Chapter 6

A Lafita & A Bateman. Modelling structural rearrangements in proteins using Euclidean distance matrices. *F1000 Research*. 9, 728 (2020). https://doi.org/10.12688/f1000research.25235.1

Chapter 1

Introduction

> *"And then, even "old school" biologists will view computational biologists as one of their own."*
>
> - Markowetz (2017): “All biology is computational biology”

In this introductory chapter, I review the scientific context of this work, including previous studies and key scientific concepts. I also describe the nature of the research project, the aims and objectives and the methodology used, and I present how this thesis document is structured into its different chapters.

1.1 Overview

Computational thinking is essential to understand the complexity of biological systems. Through computational methods, we can integrate biological knowledge and experimental data to systematically and quantitatively classify, explore and discover new biology. Computational biology enables us to see the big picture and generalise our ideas and hypotheses by systematically testing them on large sets of data, providing us with a reference map to guide future research projects in biology (Markowetz, 2017).

At the time of this work, the exponential increase in the availability of biological data is being accompanied by an equal increase in the popularity of computational approaches in biology. It is now feasible to do things that were not even imaginable 10 years ago, and new biological models and theories can now be tested solely relying on public databases.

From the start of this project, the UniProt database of protein sequences (Bateman, 2019) has tripled in size, from 70 million in 2017 to over 210 million sequences in 2021, while new protein resources derived from metagenomics already count their sequences in the billions (Mitchell *et al.*, 2018). This dramatic increase in genomic and derived protein sequences empowers us to do studies of genome and protein evolution at higher resolutions, at the individual organism rather than at the species level.

The data increase in databases of macromolecular structures, such as the Protein Data Bank (Berman *et al.*, 2000), have also greatly improved our ability to study molecular mechanisms at atomic resolutions, with new techniques such as Cryo-Electron Microscopy (CryoEM) that allow us to see bigger and more complex protein machines in action. The combination of sequence and structural data availability has also improved our ability to model the structures of novel proteins from their sequence, using evolutionary and geometrical patterns learned from protein sequences and their experimental structures to guide predictions.

Two different computational approaches are taken in this thesis to study a special subset of proteins. The first involves the use of computational techniques to explore the protein sequence space in order to discover and classify new proteins of interest. The second involves translating empirical observations to general principles and computational models in order to gain biological insights and make predictions about proteins.

1.2 Background

Most of the complexity that can be observed in biological systems emerges from their molecular structure. Over the course of evolution, new molecular components with potentially novel functions arise through duplication, recombination and diversification of existing elements (Jacob, 1977; Chothia *et al.*, 2003) These components, or building blocks, are found across different molecular levels: from individual amino acids and nucleotides that form proteins and DNA, to protein subunits that assemble into higher order symmetric complexes, such as viral capsids.

Domains are the evolutionary units of proteins, generally associated with independent and well-defined functions (such as cell localization, protein and DNA binding, and enzymatic and catalytic activities) and that fold into conserved globular structures (Hubbard *et al.*, 1999). Protein domains are classified into distinct evolutionary-related families, and collected in databases such as Pfam for protein sequences (El-Gebali *et al.*, 2019), and ECOD for structures (Cheng *et al.*, 2014). Throughout evolution, these individual domains are combined to form multidomain proteins with diverse architectures and complex functions (Vogel *et al.*, 2004; Bashton and Chothia, 2007; Levitt, 2009).

Repetitions of sequence and structure are also ubiquitous in molecular elements and their study is of vital importance to understand function and evolution of biological systems. For example, short tandem repeats in DNA have been associated with bacterial adaptation mechanisms (Moxon *et al.*, 2006; Zhou *et al.*, 2014), and many protein domain folds have originated via internal repetitions (Andrade *et al.*, 2001). Repeating elements in proteins have been extensively studied in the past, and a large number of detection methods developed to search for sequence and structural repetitions in proteins. Protein repeat types have been further classified in databases such as RepeatsDB (Paladin *et al.*, 2017).

The main focus of this work is the study of tandem domain repeats, an understudied protein domain architecture that consists in sequential homologous domains repeated in tandem. These tandem domain repeats — also known as "beads on a string" in RepeatsDB for their structural appearance — correspond to a separate category of repeat proteins, where repeating units are independently folding domains rather than short structural motifs. I will use the term "tandem domain repeats" to refer to this type of repeat protein throughout this thesis, not

to be confused with "tandem repeat domains", a term used to refer to single domains formed by repetition of short structural motifs such as Ankyrin, Armadillo or Leucine-rich repeats. Previous studies have estimated that 11% of proteins in multicellular organisms contain these tandem domain repeats (17% in humans). This fraction is lower in Prokaryotes (4%), partly due to their smaller overall fraction of multidomain proteins (Apic *et al.*, 2001; Björklund *et al.*, 2006).

Several biological functions have been associated to the formation of tandem domain repeats, such as protein complex assembly, cell-adhesion and cell-signaling (Han *et al.*, 2007). For example, the intersarcomeric filament titin is the largest protein known to date and contains over one hundred tandem immunoglobulin domains, which confer passive elasticity to striated muscles (Figure 1.1A). The Early Growth Response protein 1 (EGR1) is formed by three tandem Zinc finger domains that together dictate its binding affinity and specificity to a sequence of double stranded DNA (Miller *et al.*, 1985) (Figure 1.1B). Other proteins with regulatory and signaling functions, like human ubiquitin, are expressed as long polypeptide chains of identical tandem repeats that are post-translationally cleaved into single-domain proteins, facilitating their rapid expression and release in large quantities (Figure 1.1C). In addition to their biological function, the multivalency associated with the presence of repeated units in a polypeptide chain has been associated to unusual biophysical properties in these proteins, such as liquid-gel phase transitions (P. Li *et al.*, 2012).

Many bacterial surface proteins, such as the surface protein G (SasG) from *Staphylococcus aureus* (Gruszka *et al.*, 2012), contain nearly identical tandem domain repeats, also known as "stalk" domains, expected to be rare in nature because protein sequences tend to diversify over the course of evolution. These proteins have been described as "fibrillar adhesins" in the literature because they form thin fibrils at the bacterial cell surface visible with Electron Microscopy techniques (Back *et al.*, 2020). Stalk domain repeats have been the focus of several microbiology studies (Wästfelt *et al.*, 1996; Roche *et al.*, 2003), but their function remained elusive until structures revealed they fold into stable globular domains (Gruszka *et al.*, 2012; Whelan *et al.*, 2019). Stalk domains form rigid rod-like structures that project proteins out of the bacterial surface and enable important functions such as cell-adhesion and biofilm formation (Gravekamp *et al.*, 1996; Corrigan *et al.*, 2007).

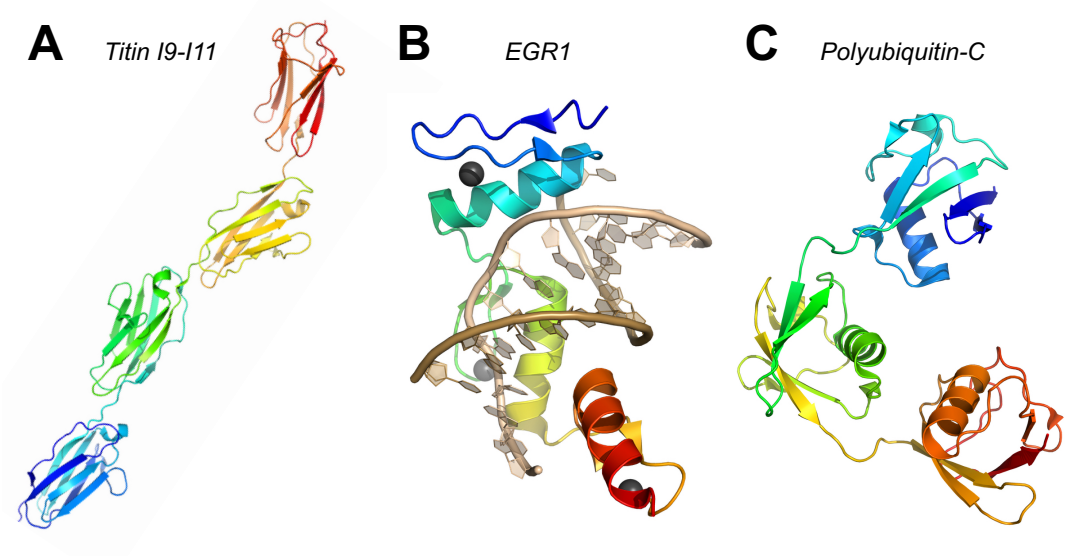

FIGURE 1.1 Three examples of human proteins containing tandem domain repeats. A) Four tandem immunoglobulin domains (Pfam:PF07679) of the intrasarcomeric filament titin share 35-40% sequence identity (PDB:3B43); B) the binding of the EGR1 protein to the double stranded DNA is mediated by three tandem Zinc finger domains (Pfam:PF00096) with 45-60% sequence identity (PDB:4R2A); and C) two human genes, UBB and UBC, contain multiple identical ubiquitin domains (Pfam:PF00240) in tandem (PDB:5H07). Protein chains are colored by sequence from blue (N-terminal) to red (C-terminal). Structures visualised using PyMol.

Protein domains generally fold independently (Batey *et al.*, 2008), but interactions between adjacent domains in multidomain proteins, shown to occur experimentally (Han *et al.*, 2007), can lead to misfolding and aggregation of proteins, causing major problems for cells and organisms. Large multidomain proteins are therefore expected to be more prone to these misfolding and aggregation effects due to their complicated folding energy landscapes; large arrays of tandem domain repeats represent an extreme case. In a seminal study, Wright *et al.* (2005) reported a direct correspondence between the sequence identity of adjacent domains and their aggregation rate: aggregation was the highest in pairs of identical domains, and very high in pairs of adjacent domains over 70% sequence identity. Further single-molecule fluorescence studies by M. B. Borgia *et al.* (2011) and A. Borgia *et al.* (2015) identified specific misfolded conformations, caused by the formation of native-like interactions between adjacent domains that were strongest for identical domain pairs.

These observations led to preliminary bioinformatics analyses on two domain families — Immunoglobulin (Pfam:PF07679) and Fibronectin type III (Pfam:PF00041) — that initially revealed overall low sequence identity (rarely over 30%) in pairs of adjacent tandem domain repeats (Wright *et al.*, 2005). The authors suggested the presence of positive selection to rapidly diversify sequences of adjacent domain repeats, presumably to avoid misfolding. Although this might be true for some protein domain families (or the majority of them), adjacent domains with high sequence identity do exist in natural proteins, such as the ones observed in bacterial surface proteins, and might also be common in other protein families.

Despite many studies on protein repeats can be found in the scientific literature, much of the emphasis has been put on short structural motifs that generate domain units, so-called "tandem repeat proteins" (Andrade *et al.*, 2001). Tandem domain repeats have been traditionally considered a property of protein domain architectures and remain largely understudied; there have been relatively few comprehensive studies of tandem domain repeats in proteins, which are over ten years old at the time of this work and have focused on the identification of domains from pre-existing domain families (Apic *et al.*, 2001; Björklund *et al.*, 2006). The exponential increase in protein sequence and structural databases offers now an opportunity to better understand the origin, evolution and function

of tandem domain repeats in natural proteins, and to explore domain repeats from uncharacterised families.

Furthermore, little is known about nearly identical tandem domain repeats. Previous studies did not consider the sequence identity of adjacent domains, but experimental observations by Wright *et al.* (2005) suggest that highly similar adjacent domains in proteins would be susceptible to misfolding and aggregation. Natural proteins with highly similar tandem domain repeats, such as stalk domains in bacterial surface proteins, challenge these experimental results and open new research questions about the folding and aggregation of multidomain proteins. Biochemical and computational studies have suggested that native-like interactions between adjacent domains are the main cause of misfolding (M. B. Borgia *et al.*, 2011; Tian and Best, 2016), but these studies are limited to small subsets of domains and general misfolding determinants in tandem domain repeats are still unknown.

1.3 Aims and objectives

The primary research hypothesis is that tandem domain repeats, specially those with high sequence similarity, create misfolding and aggregation challenges for proteins. Tandem domain repeats found in natural proteins are therefore expected to be subject to unique evolutionary and selective pressures, and potentially evolved mechanisms to avoid protein misfolding and aggregation.

In this study, I aim to understand the role of tandem domain repeats in natural proteins, and to discover unique sequence and structural properties potentially related to protein misfolding and aggregation. The research project has the following objectives:

1. Survey the prevalence and distribution of tandem domain repeats across organisms and protein families, and improve their coverage and classification in Pfam.
2. Investigate the emergence and evolution of tandem domain repeat regions, characterising the size, frequency and location of domain repeat expansions; and their impact in protein function.
3. Select tandem domain repeat candidates for further biochemical and structural studies, in collaboration with experimental research groups.

4. Identify unique sequence and structural properties of tandem domain repeats, focusing on potential misfolding and aggregation resistance mechanisms.
5. Understand potential misfolding mechanisms caused by adjacent domain interactions using molecular modelling techniques.

To our knowledge, this work constitutes the most comprehensive study of tandem domain repeats in proteins, and the first one to focus on nearly identical tandem domain repeats. It represents a step towards understanding how tandem domain repeats are used in natural proteins, with important implications for the engineering and design of multidomain proteins.

1.4 Methodology

In this thesis, the reader will find a research approach based on bioinformatics and other computational analyses. I use a wide range of public databases: from genomics data to protein sequence and structure resources, such as UniProt, Pfam and the Protein Data Bank (PDB). I also make use of smaller and more specialised datasets, such as subsets of bacterial genomes sequenced with high quality long-read technology, and other primary data directly from our experimental collaborators. One of the key challenges of this project has been to combine large heterogeneous datasets into cohesive and comprehensive results.

I have used openly available bioinformatics tools for a wide range of applications, from searching for protein homologs in sequence databases to detecting tandem sequence repeats. I have also written scripts to calculate custom sequence and structural properties of proteins and genes, and developed two new computational tools that fill existing scientific software gaps in modelling the structure and energetics of proteins.

From the start, this project has been highly collaborative; results presented throughout this thesis reflect mostly the bioinformatics side of the project, but many of the scientific insights described have come by integrating computational and experimental results.

1.5 Thesis structure

This thesis is structured with an introduction (this chapter), five results chapters, and a final conclusions chapter. Each results chapter is written as a self-contained piece of work, with its own introduction, results, methods and discussion sections. Methods and results are interspersed within different chapter sections to prioritise the flow of reading, but a Methods section with extended descriptions and methodological details is included at the end of each chapter.

Chapter 1 In this first introductory chapter, I have described the scientific context, the nature of the research project, and defined key scientific concepts. I also presented the research aims and methodology in general terms.

Chapter 2 In the second chapter, I describe two approaches to detect tandem domain repeats in proteins, based on repeat detection methods and Pfam domain families. I present the prevalence and distribution of tandem domain repeats across proteins, organisms and domain families, and identify characteristic properties of highly similar tandem domain repeats.

Chapter 3 The third chapter of the thesis is focused on Periscope proteins, a new class of bacterial surface proteins implicated in biofilm formation and cell adhesion. I describe a pipeline to identify Periscope proteins in bacterial genomes and study their domain repeat number variability across bacterial strains.

Chapter 4 In the fourth chapter, I present our efforts to experimentally determine and analyse structures of tandem domain repeats, in collaboration with structural biology groups. I describe tandem domain repeat structures, highlighting cases of structural malleability, and a structural determination pipeline for a subset of tandem domain repeats selected from Pfam families.

Chapter 5 In this chapter, I develop a new computational method to systematically estimate the stability of tandem domain swaps, a misfolded conformation observed experimentally in highly similar tandem domain repeats. I use the method to calculate misfolding propensities across domain structures in the ECOD database and report structural determinants of misfolding.

Chapter 6 In the last results chapter, I present a novel method to model protein conformations using their distance matrices that is well-suited to study structural rearrangements such as domain swapping and domain atrophy, and I explore further applications of these models to investigate protein geometry and flexibility.

Chapter 7 In the conclusions chapter, I summarise the key findings and limitations of this study, and discuss potential implications and future research.

Chapter 2

Survey of tandem domain repeats in proteins

> *"HMMER3 is substantially more sensitive and 100- to 1000-fold faster than HMMER2. HMMER3 is now about as fast as BLAST for protein searches."*
> - Eddy (2011): "Accelerated profile HMM searches"

In this chapter, I explore the prevalence and distribution of tandem domain repeats in natural proteins. I review computational methods for the detection of tandem repeats in biological sequences and use two different approaches to identify tandem domain repeats across large protein databases. I further find several characteristic properties of tandem domain repeats by comparing their sequences to isolated domains within Pfam families.

New Pfam families for tandem domain repeats were built together by Alex Bateman and I. Results on bacterial stalk domains presented in section 2.3.2 are part of a recent study led by Vivian Monzon and submitted to a journal (Monzon, Lafita, and Bateman, 2020). I identified the stalk domain families from a dataset of bacterial fibrillar adhesins compiled by Vivian, and contributed to manuscript writing.

2.1 Introduction

Sequence and structural repeats are ubiquitous in proteins. Present-day domain folds have likely arisen through ancient repetitions of short structural motifs (Lupas *et al.*, 2001), and repeats also play important roles in biological functions such as protein and DNA binding (Andrade *et al.*, 2001). In multidomain proteins, repetitions of domains — structural and functional units of proteins — are essential for the generation of diversified and complex functions.

Two key studies by Apic *et al.* (2001) and Björklund *et al.* (2006) looked at tandem domain repeats in diverse sets of proteomes across the tree of life. Both studies concluded that tandem domain repeat regions are more prevalent (11% on average) and longer in multicellular organisms than in unicellular organisms (4–6%). These results were expected and can be in part explained by the higher overall fraction of multidomain proteins in multicellular organisms.

Apic *et al.* (2001) further found that domains in tandem repeats accounted for 24% of domain families in multicellular organisms and 10–14% in unicellular organisms. Domain families most commonly found in tandem repeat regions included Immunoglobulins, a widespread β-sandwich fold, Spectrin repeats, domains part of a cytoskeletal protein with a three-helix bundle fold, and Zinc fingers, small domains with an α-helix and two β-strands stabilized by the coordination of one or more Zinc ions that commonly bind to DNA.

The evolution of tandem domain repeat regions in proteins is driven by internal domain duplications, leading to repeat number expansions. Björklund *et al.* (2006) further studied patterns of domain duplications in tandem domain repeat regions of proteins using the pairwise sequence similarity between domains. They showed that domain repeat expansions occur more frequently in the middle of tandem repeat regions, contrary to other studies of multidomain proteins, which found that new domains are mainly added at the protein termini (Björklund *et al.*, 2005; Buljan and Bateman, 2009). Using autocorrelation vectors of domain similarity, they further found that duplication events often involve several domains at a time and that the number varies across domain families.

The work by Apic *et al.* (2001) and Björklund *et al.* (2006) set the foundations to understand the prevalence, evolution and role of tandem domain repeats in proteins. Both studies took a proteome-centric approach, manually selecting proteins from a reduced subset of up to forty well-studied organisms across the

tree of life, and considered only domains classified in known families. Databases of protein sequences and domain families have experienced a rapid increase in recent years, but there have not been any other attempts to systematically study tandem domain repeats in proteins. Besides, the coverage of tandem domain repeats by known domain families is still unknown and requires a different unbiased approach to detect tandem domain repeats without relying on pre-existing domain definitions.

Finally, none of these studies systematically analysed the sequence similarity in tandem domain repeat regions, despite its suggested role in protein folding and aggregation. Experimental observations by Wright *et al.* (2005) indicated that adjacent domains over 70% sequence identity were prone to misfolding and aggregation, and they reported that the sequence similarity between adjacent domains in two families — Immunoglobulin (Pfam:PF07679) and Fibronectin type III (Pfam:PF00041) — is rarely over 30% sequence identity and that it is on average lower than the sequence similarity between non-adjacent domains. These results were interpreted as evidence for a negative selection against highly similar tandem domain repeats, prompting a rapid diversification of adjacent domain sequences. However, highly similar tandem domain repeats do exist in natural proteins and studying their prevalence systematically could reveal more insights on their relation to protein misfolding and aggregation.

In this chapter, I explore the prevalence and distribution of tandem domain repeats in natural proteins, with a special focus on repeats with high sequence identity. In the next two introductory subsections, I review methods for the detection of tandem repeats in sequences and the classification of protein domains into evolutionarily related families, which form the basis of the two approaches I use to find tandem domain repeats in proteins. In the third subsection, I review different types of composition biases and other low complexity regions in proteins, and explore their relevance to tandem domain repeats.

2.1.1 Tandem repeat detection

Repeat detection is a common and well-studied bioinformatics problem, closely related to the alignment and comparison of biological sequences (Pellegrini, 2015). It consists in searching for two or more similar subsequences within a protein or DNA string, with the goal to find the number, location and similar-

ity of repeated segments, commonly in the form of a multiple sequence alignment. Methods with a wide variety of approaches have been developed over the years and can be grouped into methods that use homology to known repeating elements and methods that directly detect sub-optimal alignments from the self-alignment matrix, such as RADAR (Heger and Holm, 2000).

Tandem repeats represent a special case where repeating elements are immediately adjacent to each other, or within a short sequence separation, and in sequential order rather than in scattered locations across the sequence. They can be easily spotted in self dot-plots — graphical representations of pairwise sequence similarity in the form of matrices — as equispaced diagonal lines enclosed in a square subregion (Figure 2.1). Even though general repeat detection methods are applicable to tandem repeats, methods specific for the detection of tandem repeats have been developed, as they represent a simpler subproblem and more efficient algorithms can be designed.

The systematic search of tandem repeats in sequences presents several challenges, such as defining the repeat region, boundaries, and number; and several parameters have to be taken into consideration, including repeat divergence, length and number. The sequence similarity between repeats is one of the most important factors. Finding nearly identical repeats is an easier but infrequent case: most sequence repeats are divergent, and sequence diversity needs therefore to be taken into account. The repeat sequence similarity baseline determines the assumptions and heuristics that can be used, and ultimately what detection approaches are more suitable. Another important factor is the length of the repeating units. Finding stretches of short sequence repeats, such as single or dinuclotide repeats in DNA, is a simpler task than finding large tandem repeats of any length. Tandem domain repeats are long repeats of 50-200 amino acids, equivalent to 150-600 nucleotides in DNA.

Several tandem repeat detection methods have been developed. They can be broadly classified in two classes: combinatorial-based and heuristic-based methods (Lim *et al.*, 2013). Combinatorial methods exhaustively compare every possible subsequence, up to a specified length, against the full sequence to find regions that fulfill the similarity threshold. This brute-force approach is feasible only for the detection of short and nearly identical repeats, since its computational complexity is cubic on the sequence length and the maximum repeat length;

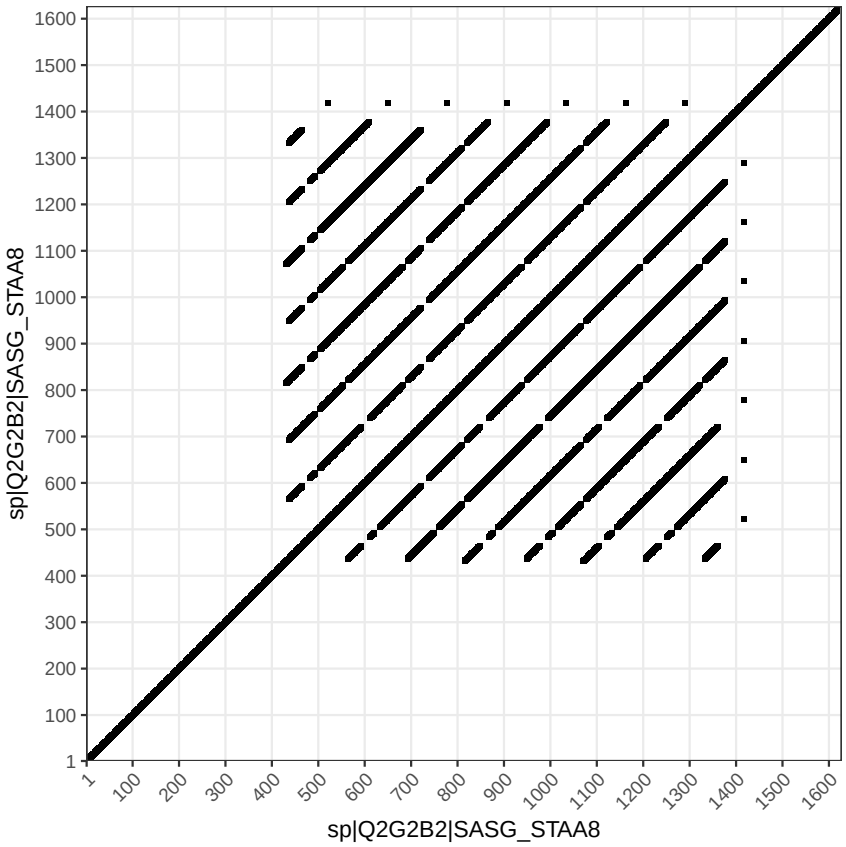

FIGURE 2.1 Self dot-plot of the SasG protein sequence (UniProt:Q2G2B2). Two identical copies of the sequence are arranged as rows and columns of the matrix and compared elementwise against each other. Entries in the matrix are colored in black for regions of identical 4-residue k-mers. Nearly identical tandem repeats of 128 residues can be observed in the middle of the protein, between positions 400 and 1400.

these methods are only applicable for short DNA repeats such as microsatellites. Heuristic-based methods overcome this limitation at the expense of reporting a less comprehensive set of repeats. Broadly, methods use small windows to scan the sequence searching for short equispaced perfect repeats, before merging them into longer repeating segments. Different heuristics have been designed for specific biological applications, typically allowing for longer and imperfect repeats up to a lower bound of repeat similarity; and they are better suited for the purposes of finding tandem domain repeats.

The method T-REKS (Jorda and Kajava, 2009) uses a very fast heuristic algorithm, which is based on the analysis of the distribution of identical short strings within the sequence using K-means clustering. It is suitable for large scale analysis and for finding large repeats with sequence identities of at least 70%, although it is optimal for repeats with high sequence identities above 90%. Although it can, in principle, handle protein and nucleotide sequences, its heuristic parameters are optimised for the analysis of protein sequences. The Tandem Repeat Finder (TRF) method (Benson, 1999) is another heuristics-based method that uses statistical criteria to find candidate tandem repeats of any size and up to 20% sequence divergence in DNA. Implementations for these two methods are openly and freely available.

2.1.2 Protein domain classification

Protein domains are classified into evolutionary-related families of similar structure and closely related functions. Pfam is a database of profile Hidden Markov Models (HMMs) — probabilistic models that capture position-specific information of sequence evolution — derived from manually curated multiple sequence alignments of related and non-redundant proteins, known as the SEED alignment (Sonnhammer *et al.*, 1998). Each SEED alignment and associated profile HMM describes a family of homologous proteins, domains or motifs, and can be used to search for related sequences across large datasets using the HMMER tool (Eddy, 2011).

Pfam classifies families into different types: family, domain, motif, repeat, coiled-coil and disordered (El-Gebali *et al.*, 2019). Repeat families are short sequence or structural units found multiple times in proteins, while domain families correspond to single globular domains and are usually associated to known

structures. Individual Pfam families are further grouped into clans of remotely related families, usually with known structural homology but without detectable sequence similarity.

Detecting tandem domain repeats in proteins is a much easier task if the repeating domains are part of a known family. Previous studies by Apic *et al.* (2001) and Björklund *et al.* (2006) have exclusively used this approach, which further allows the study of domain properties in their evolutionary context provided by other domains in the family.

2.1.3 Globular domain topologies

Protein chains fold into three-dimensional structures that show hierarchical patterns of organisation: sequential amino acid segments adopt secondary structures (α-helices and β-strands), which arrange into structural motifs (such as β-hairpins or helix-turn-helix) and form a single or multiple discrete globular units of structure, known as protein domains. The size of globular domains in proteins mostly varies between 50 and 200 residues (Xu and Nussinov, 1998).

Protein domains are further classified structurally in databases such as SCOP (Hubbard *et al.*, 1999), CATH (Sillitoe *et al.*, 2015) and ECOD (Cheng *et al.*, 2014). Structural information is richer and more conserved than sequences, allowing improved assessments of remote homology between domains and their classification into a hierarchy with several layers. At the lowest level, profile HMMs are used to classify domains into individual families with detectable sequence homology. Other higher levels use structural scores and comparisons, that sometimes require manual curation, to classify domains into different secondary structure classes, architectures and topologies.

For example, at the top level CATH classifies domains into few secondary structure classes (mainly alpha, mainly beta, alpha beta, etc), followed by architectures (up-down alpha bundle, beta barrel, beta sandwich, etc), topologies (Immunoglobulin-like, TIM barrels, Rossman folds, etc) and homologous domain families at the lowest level. The ECOD database follows a similar classification structure, although some category names differ, and it is kept up-to date with the latest release of the Protein Data Bank (PDB) (Berman *et al.*, 2000), achieving the highest classification coverage among domain structure databases.

2.1.4 Protein composition bias

Biases in amino acid composition and other low complexity regions are also common in proteins, present in up to 20% of Eukaryotic proteomes, and have been associated with several biological functions (Mier *et al.*, 2020). Low complexity regions are often the result of sequence and structural periodicity, and generally function as flexible disordered polypeptide chains without stable folded 3D structures.

Biases for certain amino acids are commonly found in proteins from extremophile organisms as a response to the impact of unusual temperature, pH and salinity conditions on protein stability and activity (Reed *et al.*, 2013). Even though most of these biased proteins fold into stable structures, they tend to be more frequently mispredicted as disordered by sequence-based tools such as IUPred (Dosztányi *et al.*, 2005), because low complexity is a characteristic of disordered proteins (Pancsa *et al.*, 2019). Other weaker types of amino acid biases are found in membrane proteins (Deber *et al.*, 1986), in certain structural motifs such as Leucine-rich repeats (LRRs) (Kobe and Deisenhofer, 1994), and in amyloidogenic proteins linked to prion phenomena and other diseases in humans, which contain Glutamine (Q) and/or Asparagine (N) rich regions (Harrison and Gerstein, 2003).

Since the beginning of the project, I have been intrigued by striking biases of amino acid composition observed in some tandem domain repeats, causing the regions to be mispredicted as disordered by IUPred, such as in the tandem G5 and E domains of the SasG bacterial surface protein. Gruszka *et al.* (2016) found that the E domain is indeed disordered in isolation, but folds into a stable structure cooperatively to adjacent G5 domains. In the context of previous studies that reported an increased aggregation propensity of highly similar tandem domain repeats (Wright *et al.*, 2005), evolutionary adaptations in the form of amino acid biases, similar to those in extremophiles, could be expected in tandem domain repeats to avoid misfolding and aggregation. In this chapter, I would like to understand the amino acid composition and other properties of tandem domain repeats, focusing on how universal and widespread they are across domain families.

Methods to evaluate low complexity and composition biases in protein sequences consist in comparing expected and observed amino acid fractions in slid-

ing windows of a fixed size along the sequence. A simple and popular metric used in the literature is the information content, in the form of Shannon entropy. Other more sophisticated methods and measures have been developed for specific types of composition biases (Jarnot *et al.*, 2020), such as the Lowest-Probability Subsequences (LPS) method (Harrison and Gerstein, 2003) to find single and/or multiple amino acid biased regions in proteins.

Amino acid composition biases are commonly related to the properties of their side-chains, such as their hydrophobicity and charge. In terms of the effect of composition bias for protein stability, several side-chain properties have been suggested to be important, such as the side-chain entropy: a measure of the degrees of freedom of amino acid side-chains (Doig and Sternberg, 1995).

2.2 *De novo* tandem domain repeat detection

The first approach that I used to find tandem domain repeats consisted in using a *de novo* tandem repeat detection method (T-REKS) to search for domain-size (50-200 residues) tandem repeats in protein sequences. I decided to focus on detecting highly similar tandem domain repeats (over 90% sequence identity) in order to find the most extreme repeat cases and mitigate the methodological limitations of *de novo* repeat detection methods.

2.2.1 Prevalence of tandem domain repeats

First, I ran the T-REKS tool across 160 million protein sequences in UniProt (version 2019_06) using a 90% repeat sequence identity threshold (details in the Methods section), and further filtered the results for domain-size repeats (length between 50 and 200 residues). In total, T-REKS detected a little over 100 thousand proteins with tandem domain-size repeats: only 0.07% of UniProt (Table 2.1). Repeats are enriched three-fold in Eukaryotes compared to Bacteria, in accordance with previous studies of multidomain protein architectures.

2.2.2 Pfam coverage of tandem domain repeats

Next, I explored the coverage of tandem domain-size repeat regions identified using T-REKS by protein families in the Pfam database. Overall, the coverage

TABLE 2.1 Prevalence of tandem domain-size repeats identified by T-REKS in UniProt. Total number of proteins with highly similar (90%) tandem domain-size repeats (# Repeats), and percentage of proteins with repeats relative to UniProt proteins for each Superkingdom (% Repeats).

	# Repeats	% Repeats
Eukaryota	65,430	0.12%
Bacteria	36,266	0.04%
Archaea	860	0.02%
Viruses	529	0.01%
Unclassified	728	0.01%
Total	103,813	0.07%

in version 33.1 of Pfam is lower (around 50%) than the 70% average sequence coverage in Pfam (El-Gebali *et al.*, 2019), and the biggest coverage gap is for repeats predicted as disordered by IUPred (Figure 2.2). Although IUPred disorder predictions might suggest that these repeats do not actually fold into globular structures and should therefore not be considered domains, it is important to note that some known globular tandem domain repeats tend to be mispredicted as disordered.

Large tandem repeats in proteins have proved to be a fruitful resource to create, correct and validate domain families. More details on how Pfam families were built from repeats are described in the Methods section. Over the course of this study, Alex and I have built 34 new Pfam domain families from tandem domain repeats identified *de novo* by T-REKS, and that were not previously covered by other families in Pfam (Table 2.2). These new families mostly belong to a small subset of domain topologies, namely β-sandwich folds from the Immunoglobulin-like E-set, Ig and Transthyretin clans; β-grasp folds in the Ubiquitin clan; and domains in a new superfamily named MBG clan, also related to Immunoglobulin-like folds. Most of these new domain families are found in bacterial cell surface proteins.

The coverage of tandem domain-size repeats has increased by about 8% since the start of this project, from 42% in Pfam 31.0 (2017) to 50% in Pfam 33.1 (2020), with over 7,000 additional repeats annotated (Figure 2.2A). Some of the newest domain families created (PF19403-19408) have not yet been included in Pfam and will be part of the next Pfam 34 release.

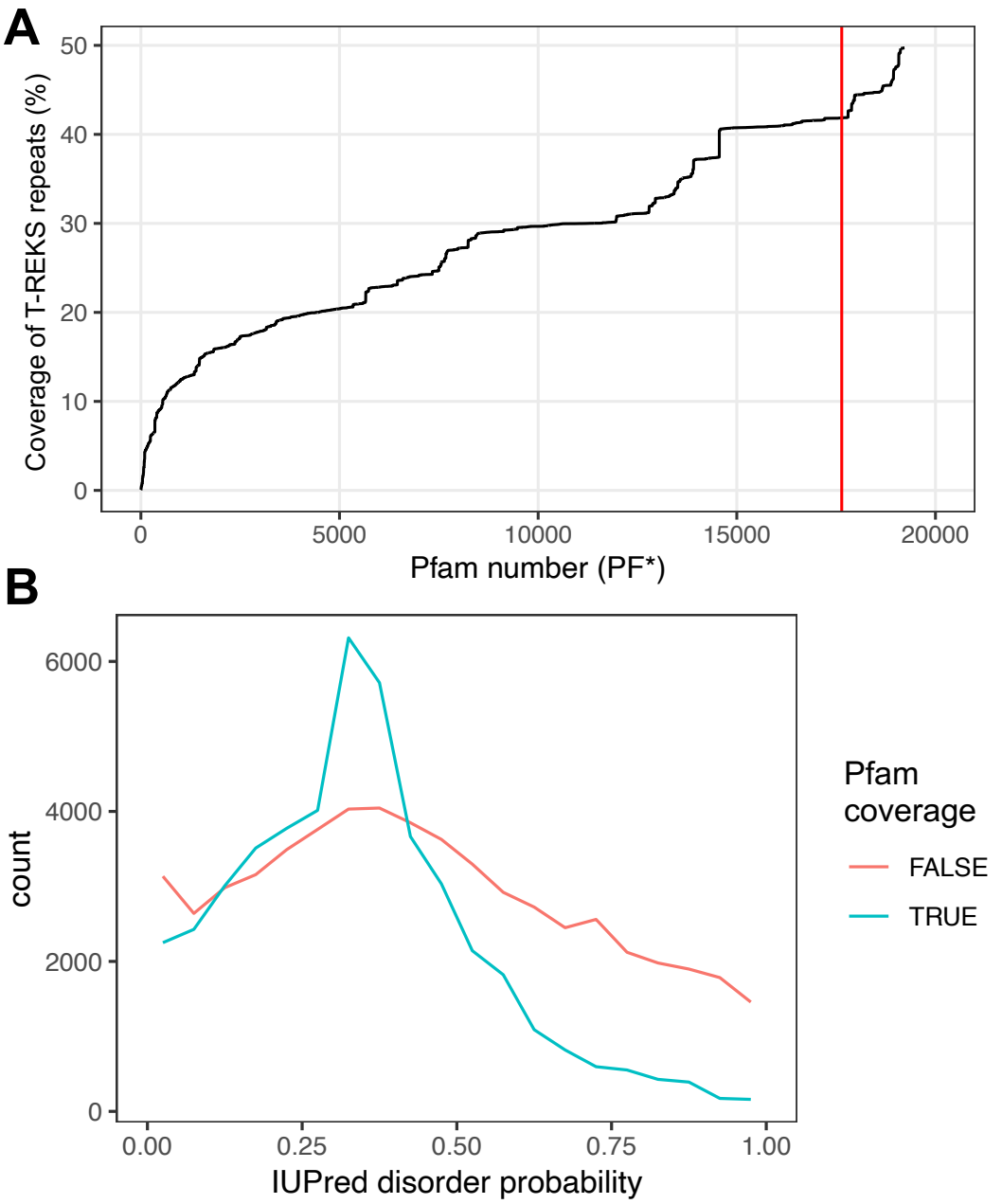

FIGURE 2.2 Coverage of highly similar tandem domain repeats by Pfam families. A) Pfam coverage of tandem domain repeat regions identified using T-REKS as a function of Pfam family number (a pseudo-time measure) up to Pfam version 33.1 (2020). Red vertical line indicates coverage in Pfam version 31.0 (2017), at the start of this project. B) Distribution of tandem domain repeat regions as a function of IUPred score (0: globular structure, 1: disordered), split in two sets: repeats covered (TRUE, blue) and not covered (FALSE, red) by Pfam.

Table 2.2 **List of new Pfam families built from tandem domain-size repeats identified by T-REKS over the course of this study (since Pfam 31.0).**

Name	Pfam ID	Pfam Clan	Clan ID
SdrD_B	PF17210	Transthyretin	CL0287
SpaA	PF17802	Transthyretin	CL0287
Cadherin_4	PF17803	E-set	CL0159
MBG	PF17883	MBG	CL0682
Cadherin_5	PF17892	E-set	CL0159
Big_6	PF17936	E-set	CL0159
Big_9	PF17963	E-set	CL0159
MucBP_2	PF17965	Ubiquitin	CL0072
Big_11	PF18200	E-set	CL0159
TQ	PF18202	NA	NA
SHIRT	PF18655	Ubiquitin	CL0072
YDG	PF18657	MBG	CL0682
MBG_2	PF18676	MBG	CL0682
QPE	PF18874	NA	NA
SSSPR-51	PF18877	Ubiquitin	CL0072
MBG_3	PF18887	MBG	CL0682
PKD_4	PF18911	E-set	CL0159
aRib	PF18938	E-set	CL0159
RibLong	PF18957	E-set	CL0159
InlK_D3	PF18981	E-set	CL0159
Flg_new_2	PF18998	Ubiquitin	CL0072
CshA_repeat	PF19076	NA	NA
Big_13	PF19077	E-set	CL0159
Big_12	PF19078	E-set	CL0159
CFSR	PF19079	NA	NA
Ig_7	PF19081	Ig	CL0011
DUF5776	PF19087	NA	NA
DUF5801	PF19116	NA	NA
SpaA_2	PF19403	Transthyretin	CL0287
DUF5977	PF19404	NA	NA
DUF5978	PF19405	NA	NA
PKD_5	PF19406	E-set	CL0159
DUF5979	PF19407	Transthyretin	CL0287
PKD_6	PF19408	E-set	CL0159

2.2.3 Structural coverage of tandem domain repeats

I further performed a T-REKS search across the Protein Data Bank (PDB) using sequence constructs directly extracted from experimental structures in the PDB, known as SEQRES, and their natural sequences from UniProt. I used SIFTS — a resource of cross-reference information between UniProt and the PDB that provides a residue by residue mapping (Velankar *et al.*, 2013) — to aggregate the two separate sets of results.

I divided structures of tandem domain repeats detected into three cases: structures with two or more domain repeats from a natural protein (complete), structures of only a single domain from a natural protein containing tandem domain repeats (partial), and structures of artificial constructs of tandem domain repeats (artificial). In total, I found 97 structures of these tandem domain repeats in the PDB (0.2% of 41 thousand unique protein chains), of which 38 correspond to artificial constructs, 47 to partial single-domain structures, and only 12 to complete tandem domain repeats (Figure 2.3A).

Natural tandem domain repeat structures found in the PDB correspond to proteins with varying numbers of tandem repeats, from two (the majority) up to 14 (Figure 2.3B). The structure with the highest number of tandem repeats covers seven repeats of the natural sequence. Most tandem domain repeats are of high sequence identity, above 90% (Figure 2.3C), and the repeat length mostly varies in the range 50-150 residues.

It was surprising to see such a large fraction (40%) of artificial constructs, many of which correspond to homomeric proteins being fused into a single chain, such as the three pilin subunits shown in Figure 2.3F. The natural hits (partial and complete) were predominantly tandem domain repeats in bacterial surface proteins, such as the B domains in Immunoglobulin G-binding protein A (Figure 2.3D), but also included domains in human proteins such as Obscurin (Figure 2.3E) and polyubiquitin, and small proteinase inhibitors and antifreeze proteins.

2.3 Pfam-based tandem domain repeat detection

The second approach I used to detect tandem domain repeats is similar to previous studies by Apic *et al.* (2001) and Björklund *et al.* (2006) and consisted in using pre-existing domain families in Pfam. I used the domain database in Pfam 31.0

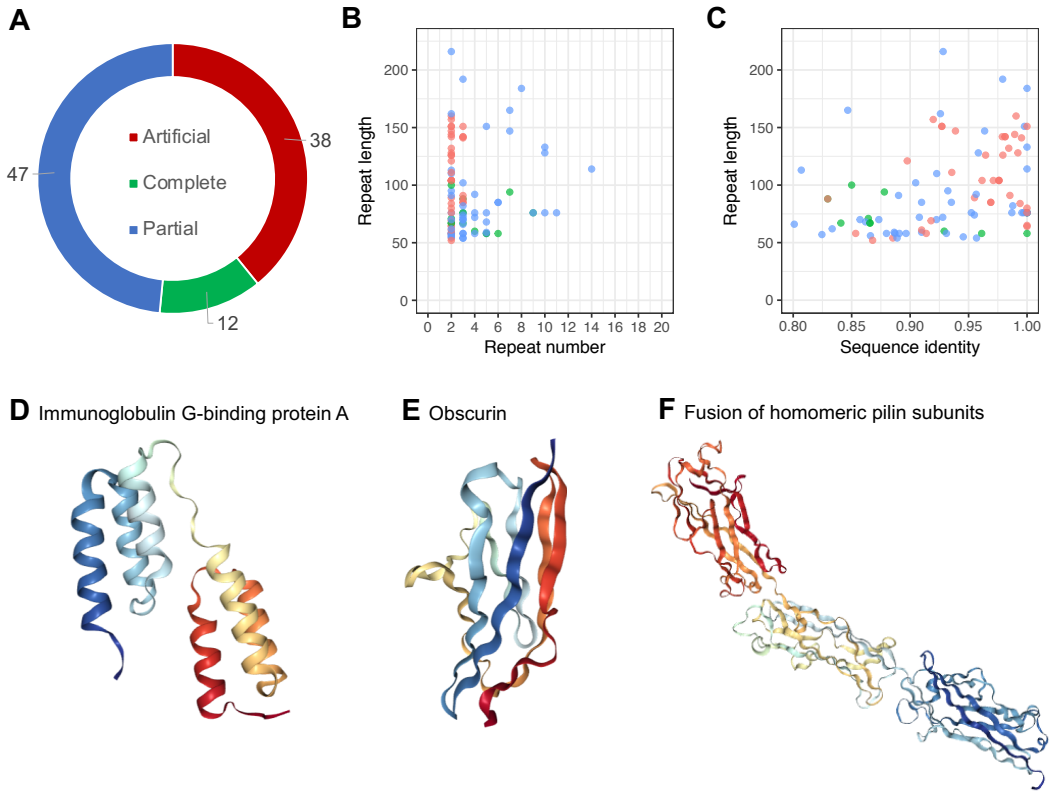

FIGURE 2.3 Structures of tandem domain repeats in the Protein Data Bank. A) Number of tandem domain repeat structures found by T-REKS in the PDB, and split by category: artificial construct, natural single-domain structure (partial) and natural multidomain structure (complete). Scatter plot of the number of repeats vs repeat length (B) and T-REKS sequence identity (C). Below, structures of three examples of tandem domain repeats in the PDB: two tandem 3-helix bundle B domains from Ig-binding protein (D), single Ig-like domain from the human Obscurin protein (E), and artificial construct of three identical pilin subunits fused in tandem (F). Structures visualised using PyMol.

(released in May 2017), which is obtained by searching the Pfam HMM model library against the UniProt Reference Proteomes using the HMMER tool. Using the sequence ranges of domain hits, I classified domains in two categories: "tandem" repeats, if two hits from the same family are adjacent and within a distance below 30 residues, and "isolated" otherwise.

I analysed 41.7 million domain hits (41,679,621 domains) from 5,130 "Domain" type Pfam families, out of which 38.9 million were isolated and 2.7 million (6.6%) were in tandem repeats. I further considered the sequence similarity between tandem domain repeats using a 70% sequence identity threshold (suggested by Wright *et al.* (2005) to be aggregation-prone), and defined a third category as high sequence identity tandem domain repeats, abbreviated as "HITRDs", if their sequence similarity is above the threshold. I found that a little over 130 thousand (133,810) tandem domain repeats are HITRDs, which is 5% of tandem and 0.3% of the total number of domains. This results show again that highly similar tandem domain repeats are rare among proteins and domains.

2.3.1 Distribution across domain families

Tandem domain repeats are prevalent (defined as more than 1% of the domains) in around seven hundred families, which is 14% of the total five thousand domain type families in Pfam (Figure 2.4A). At least one adjacent domain pair above the 70% sequence identity threshold was found in 40% of the families, but the average similarity between adjacent domains is low in most families (Figure 2.4A). There are, however, few families where tandem domains are highly similar, such as Cohesin (Pfam:PF00963), with an average 73% sequence identity. Other domain families rarely exist in isolation, like the Olduvai domain (Pfam:PF06758) with 84% of tandem domains, 45% of which share adjacent sequence identities above 70%.

Families with the highest prevalence of HITRDs (Table 2.3) include many bacterial cell surface domains (IgG_binding_B, B, Rib and GA-like), as well as other interesting proteins like the RP1-2, a structural domain found in spider silk strand proteins. In accordance with the structures of tandem domain repeats in the PDB, families for ubiquitins, NPA and small proteinase inhibitors are at the top of the list.

I further compared the sequence similarity between adjacent and non-

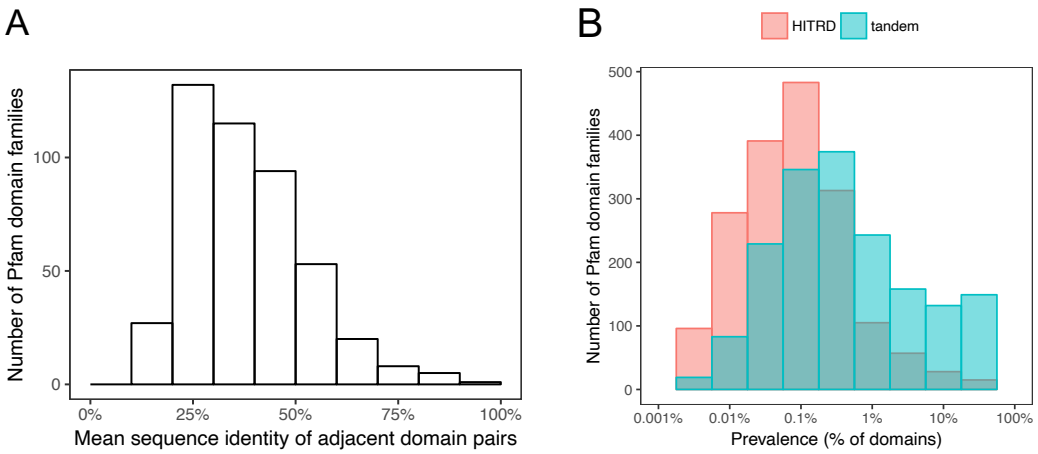

FIGURE 2.4 Prevalence of tandem domain repeats in Pfam. A) Distribution of average repeat sequence identity in adjacent domain pairs across Pfam domain families with at least three tandem domain repeats. B) Prevalence of all tandem domain repeats (tandem) and tandem domain repeats with high sequence identity above 70% (HITRDs) across Pfam domain families. Pairwise sequence identity calculated as described in the Methods section.

TABLE 2.3 Pfam domain type families with the highest prevalence of HITRDs. Top 15 Pfam domain families sorted by % of HITRDs and with more than 20 HITRDs. Pfam accession numbers (Pfam ID), names, clans, percentage of HITRDs (pHITRDs as %), and numbers of HITRDs, Tandem (excluding HITRDs) and Isolated domains for each family.

Pfam ID	Pfam name	Clan	pHITRDs	nHITRDs	nTandem	nIsolated
PF01378	IgG_binding_B	Ubiquitin	88	60	0	8
PF02216	B	B_GA	87	781	113	7
PF05386	TEP1_N	NA	75	118	38	2
PF02428	Prot_inhib_II	NA	57	354	171	91
PF12042	RP1-2	NA	52	158	38	107
PF06758	Olduvai	NA	45	590	515	205
PF08789	PBCV_basic_adap	NA	42	220	154	146
PF00299	Squash	Pept_Inhib_IE	39	36	0	57
PF14861	Antimicrobial21	NA	36	64	79	36
PF00240	Ubiquitin	Ubiquitin	31	9806	2227	19326
PF02013	CBM_10	NA	31	80	118	62
PF17573	GA-like	B_GA	28	63	22	142
PF08428	Rib	E-set	27	1133	1896	1175
PF08230	CW_7	NA	27	162	198	245
PF16469	NPA	NA	23	137	395	51

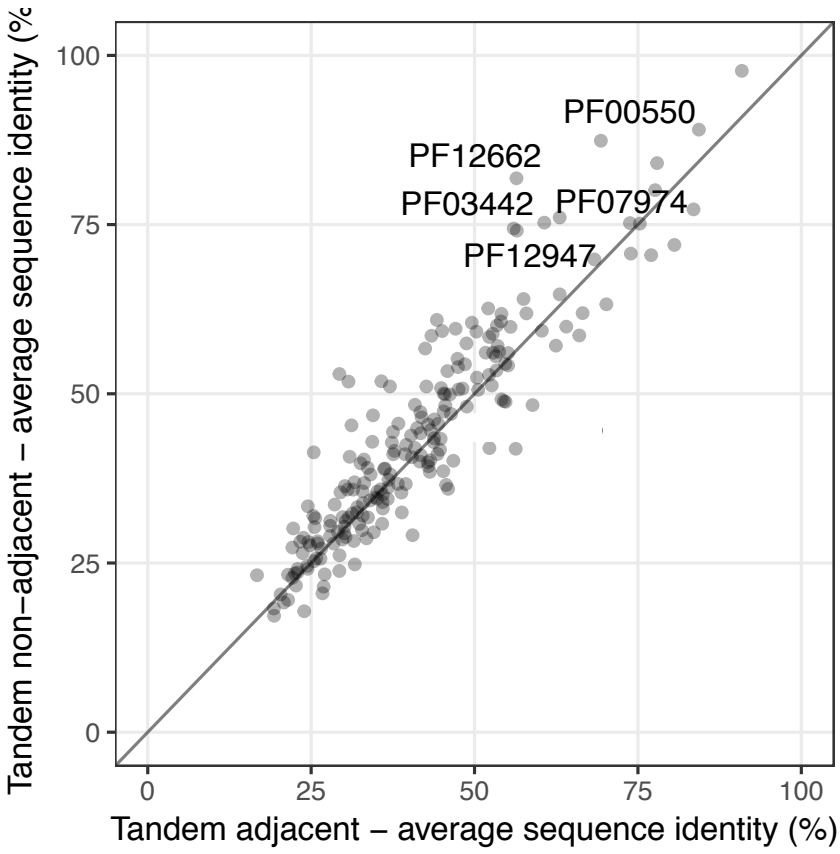

FIGURE 2.5 Comparison of the average sequence identity between tandem adjacent and non-adjacent domains in Pfam families. Few Pfam accession numbers are highlighted for families with a high average sequence identity of adjacent domains, but lower than non-adjacent domains. Pairwise sequence identity calculated as described in the Methods section.

adjacent domains in tandem domain repeat regions in all Pfam families, extending the analysis by Wright *et al.* (2005), who found a lower average sequence identity in adjacent domains compared to non-adjacent domains. I found that, for the majority of families, this observation is true, but the differences of sequence similarity between adjacent and non-adjacent are overall small. The highest differences are observed in families with high adjacent domain similarity (Figure 2.5). I also find several families where the sequence similarity of adjacent domains is higher than that of non-adjacent domains, breaking Wright *et al.* (2005) rule.

2.3.2 Bacterial stalk domain families

Despite their lower prevalence in Bacteria, many of the highest similarity and longest tandem domain repeat regions found in this study were part of bacterial cell surface proteins. In addition, our experimental collaborators were specially interested in these repetitive bacterial surface proteins, so in several analyses of this thesis I have focused on bacterial proteins. Alex and I were therefore interested in compiling a list of tandem repeat domains found in bacterial surface proteins.

Our focus has been in a special type of bacterial surface proteins, known as fibrillar adhesins, which are composed of a central repetitive region with a large number of tandem repeated domains and other domains with adhesive function commonly placed at the terminal of the protein opposed to the bacterial surface (Back *et al.*, 2020). In order to study tandem domain repeats that form stalks in bacterial fibrillar adhesins, Vivian gathered over 25 thousand bacterial proteins containing at least one hit to a set of 24 manually curated adhesive Pfam domain families extracted from known fibrillar adhesins in the literature (Monzon *et al.*, 2020).

In total, I identified 68 domain families that form tandem domain repeats associated with fibrillar adhesive domains using the Pfam-based detection approach (Table 2.4). Many of these stalk domain families had been previously created by us as part of the prevalence study, including families for the SHIRT domain (a member of the Ubiquitin clan), families in the newly created Mirror Beta-Grasp (MBG) clan and other bacterial Ig-like domains (Big) in the E-set clan. Domain families were checked manually in Pfam to ensure their description, length, species distribution and domain architecture matched the bacterial stalk domain definition.

2.3.3 Properties of tandem domain repeats

Next, I studied the properties of tandem domain repeats and compared them to other isolated domains from the same Pfam family. Over the course of the project, I have looked at several different properties. Here, I present the most relevant results.

TABLE 2.4 List of stalk domain families from bacterial fibrillar adhesins.

Name	Pfam Clan	Pfam ID	Clan ID
I-set	Ig	PF07679	CL0011
Ig_7	Ig	PF19081	CL0011
MucBP	Ubiquitin	PF06458	CL0072
Flg_new	Ubiquitin	PF09479	CL0072
MucBP_2	Ubiquitin	PF17965	CL0072
SHIRT	Ubiquitin	PF18655	CL0072
SSSPR-51	Ubiquitin	PF18877	CL0072
Flg_new_2	Ubiquitin	PF18998	CL0072
Cadherin	E-set	PF00028	CL0159
fn3	E-set	PF00041	CL0159
PKD	E-set	PF00801	CL0159
DUF11	E-set	PF01345	CL0159
TIG	E-set	PF01833	CL0159
Big_2	E-set	PF02368	CL0159
Big_1	E-set	PF02369	CL0159
HYR	E-set	PF02494	CL0159
Calx-beta	E-set	PF03160	CL0159
He_PIG	E-set	PF05345	CL0159
Big_3	E-set	PF07523	CL0159
CARDB	E-set	PF07705	CL0159
Rib	E-set	PF08428	CL0159
Big_3_2	E-set	PF12245	CL0159
Big_5	E-set	PF13205	CL0159
Cadherin_3	E-set	PF16184	CL0159
DUF5011	E-set	PF16403	CL0159
Big_3_5	E-set	PF16640	CL0159
Cadherin_4	E-set	PF17803	CL0159
Cadherin_5	E-set	PF17892	CL0159
Big_6	E-set	PF17936	CL0159
Big_9	E-set	PF17963	CL0159
Big_11	E-set	PF18200	CL0159
PKD_4	E-set	PF18911	CL0159
InlK_D3	E-set	PF18981	CL0159
Big_13	E-set	PF19077	CL0159
Big_12	E-set	PF19078	CL0159
PKD_5	E-set	PF19406	CL0159
PKD_6	E-set	PF19408	CL0159
Cna_B	Transthyretin	PF05738	CL0287
FctA	Transthyretin	PF12892	CL0287
CarboxypepD_reg	Transthyretin	PF13620	CL0287
SdrD_B	Transthyretin	PF17210	CL0287
SpaA	Transthyretin	PF17802	CL0287
SpaA_2	Transthyretin	PF19403	CL0287
DUF5979	Transthyretin	PF19407	CL0287
G5	G5	PF07501	CL0593
GA	B_GA	PF01468	CL0598
B	B_GA	PF02216	CL0598
FIVAR	B_GA	PF07554	CL0598
DUF1542	B_GA	PF07564	CL0598
MBG	MBG	PF17883	CL0682
MBG_2	MBG	PF18676	CL0682
MBG_3	MBG	PF18887	CL0682
Fn_bind	NA	PF02986	NA
SlpA	NA	PF03217	NA
SSURE	NA	PF11966	NA
Antigen_C	NA	PF16364	NA
AgI_II_C2	NA	PF17998	NA
TQ	NA	PF18202	NA
Endotoxin_C2	NA	PF18449	NA
Trp_ring	NA	PF18669	NA
QPE	NA	PF18874	NA
CshA_repeat	NA	PF19076	NA
CFSR	NA	PF19079	NA
DUF5776	NA	PF19087	NA
DUF5801	NA	PF19116	NA
DUF5977	NA	PF19404	NA
DUF5978	NA	PF19405	NA

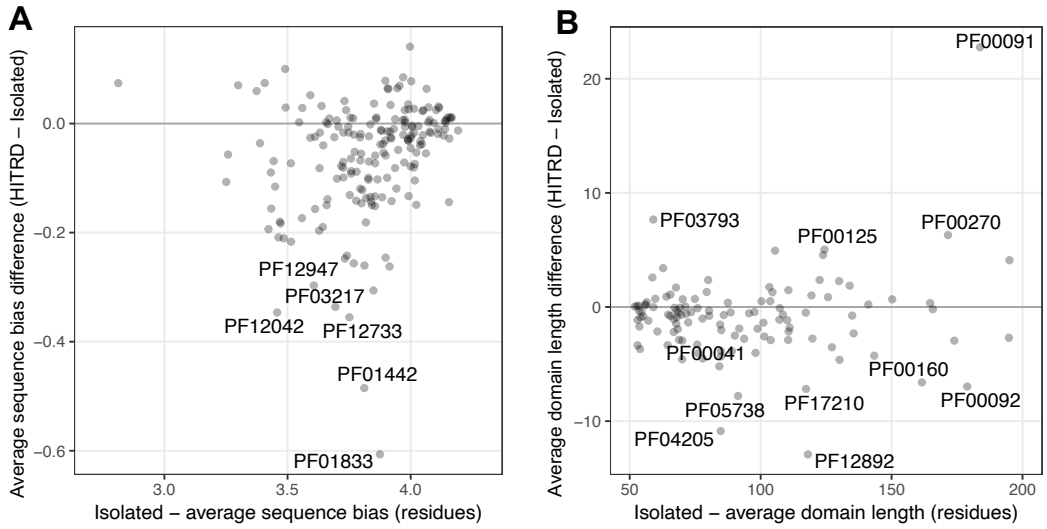

FIGURE 2.6 Comparison of the sequence composition bias and domain length between highly similar tandem domain repeats (HITRDs) and isolated domains across Pfam families. Points correspond to the average sequence bias, calculated as the absolute Shannon entropy (A), and domain length (B) for each domain family. On the x-axis, the average bias and length of isolated domains is shown; on the y-axis the difference between the average bias and length of HITRDs and isolated domains is shown, negative values indicating lower averages in HITRDs. Points below the 0 line represent domain families with a lower sequence entropy (higher bias) and shorter domain length in HITRDs than in isolated domains. Low Shannon entropy corresponds to more biased amino acid composition and lower sequence complexity. Labels shown for Pfam families with the biggest differences between HITRDs and isolated domains.

Sequence composition bias

First, I calculated the bias in amino acid composition of domains across Pfam families, comparing domains found in tandem repeats and in isolation. Alex and I observed that overall, highly similar tandem domain repeats (HITRDs) are more biased than isolated domains (Figure 2.6A).

In some families such as the TIG domains, this amino acid bias is extreme and correlates strongly with the adjacent domain similarity (Figure 2.7A). The most enriched residues are Threonine, Glycine and Alanine (Figure 2.7B), and these same enriched amino acids are common to the majority of HITRDs in the TIG family (they cluster in the sequence composition space in the PCA plot of Figure 2.7C).

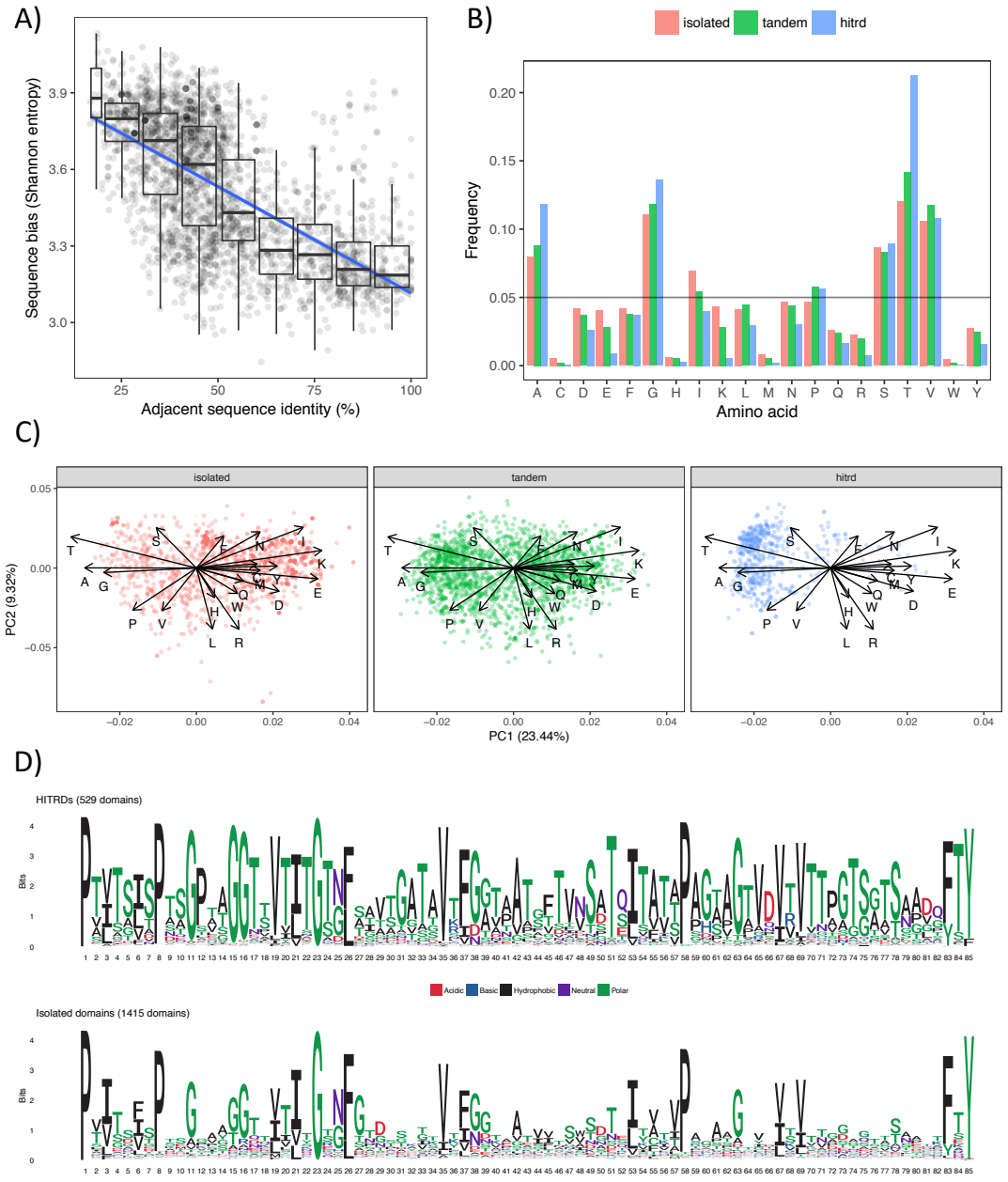

FIGURE 2.7 Analysis of the sequence composition bias in bacterial TIG domains (Pfam: PF01833). A) Scatter points and binned box-plots of the amino acid bias (Shannon entropy) as a function of the adjacent sequence identity, calculated as described in the Methods section. Linear regression fit shown as a blue line. B) Average domain frequencies of each amino acid separated by domain context: isolated, tandem (<70%) and HITRD (>70%). C) Principal Component Analysis (PCA) of the amino acid composition vectors calculated from the domain sequences in the family, split by domain context. D) Comparison of the HMM profile logos of HITRDs (top) and isolated domains (bottom). Amino acids are coloured by side-chain property.

Other Pfam families show similar amino acids biases in highly similar tandem domain repeats, such as bacterial surface domains Rib, G5, FctA and DUF1542, and other eukaryotic domain families such as NPA (additional plots for these families can be found in Appendix A: Figures A.1-A.6). For the fn3 family, a stronger correlation between the sequence bias and domain similarity is found in bacterial domains than in Eukaryotic domains, although the latter are also more biased than isolated domains (Figure A.1)

Amino acid side-chain properties

I explored several properties of the enriched amino acids in order to understand the source of the sequence bias. In some families, the side-chain entropy correlates well with the amino acid bias observed, for example in Rib, NPA, and DUF1542 (Appendix A). In these families, the average side-chain entropy of highly similar tandem domain repeats is lower, indicating a lower flexibility of side-chains that could have an impact on the domain stability and folding dynamics. For other families, however, no correlation between the bias and the side-chain entropy could be observed (for example in fn3, G5 and FctA), but no families with a significantly higher side-chain entropy in highly similar tandem domain repeats were found either.

I further explored other properties of the biased amino acids, such as hydrophobicity, charge, polarity and isoelectric point of the domain, but none showed significant correlations with the observed amino acid biases.

Domain length

Another unique property observed in highly similar tandem domain repeats is that they are on average shorter than isolated domains, and that the average domain lengths also correlate with the sequence identity between adjacent domains (Figure 2.8). This correlation is very strong in some families, such as fn3 and FctA (both Immunoglobulin-like folds), which show a combination of larger domain insertions and lower gap fractions along their HMM profile in the isolated domains compared to highly similar tandem domain repeats. Other families only show weak correlations or do not show any correlation at all, and for a minority of families the average domain length is higher in tandem domain repeats (Appendix A).

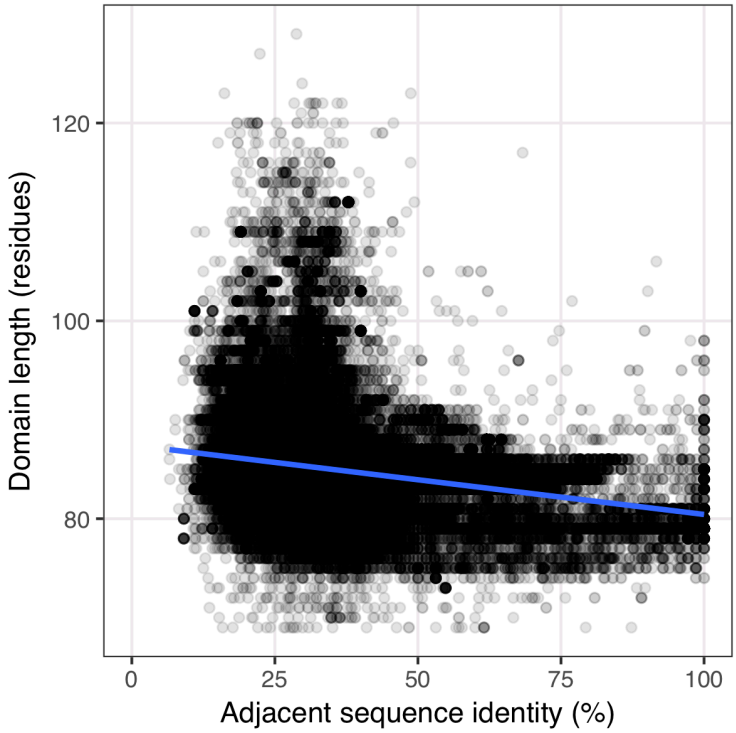

FIGURE 2.8 Correlation between the domain length and the adjacent sequence identity in fn3 domains (Pfam: PF00041). Points correspond to individual tandem domain repeats from the fn3 family. The blue line is a linear regression fit of the points. The pairwise sequence identity between domains is calculated as described in the Methods section.

2.4 Discussion

In this chapter, I conducted a search of tandem domain repeats in natural proteins using two approaches: a *de novo* identification using tandem repeat detection methods, and a domain-based identification using homology to pre-existing Pfam domain families. The unbiased nature of the *de novo* detection approach enabled a more comprehensive study of their prevalence and an improvement of the Pfam coverage of tandem domain repeats from 42% to 50% through 34 new domain families.

The prevalence survey indicates that highly similar tandem domain repeats (above 90% sequence identity) are very rare in natural proteins, around 0.12% of proteins in Eukaryotes and 0.04% in Prokaryotes, and only account for 0.3% of the total number of domains in Pfam. There might be several reasons for this low prevalence. First, proteins accumulate neutral mutations throughout their evolution that slowly diversify the sequences of tandem domains, suggesting that most tandem domain duplication events in proteins are evolutionarily remote. A second possible explanation is the presence of selective pressures that limit the type of domains that can be tandemly repeated with high sequence similarities, and that push sequence diversification through higher mutation rates. Although I did not find direct evidence for such negative selection pressures, I discovered that the general principle by Wright *et al.* (2005) — that adjacent tandem domains are less similar than non-adjacent ones — holds for the majority of Pfam families with the highest average adjacent sequence similarity, although several exceptions to the rule were found. I further observed unique properties in highly similar tandem domain repeat sequences, such as amino acid composition biases and domain length reduction, which might be directly related to strong selective pressures in these domains.

In line with previous studies by Apic *et al.* (2001) and Björklund *et al.* (2006), tandem domain repeats are found across a wide range of domain families in Pfam (14% of domain type families) including both α-helical and β-sheet folds. However, families with the highest prevalence of highly similar tandem domain repeats and new families built from T-REKS repeat sequences correspond to a limited subset of topologies, most of them related to Immunoglobulin-like, Ubiquitin-like and three-helical bundles.

In the analysis of adjacent domain sequence similarity I found several domain

families with a high average sequence identity between adjacent domains, above 50% and much higher than the 30% average previously reported by Wright *et al.* (2005). This finding suggests that the misfolding and aggregation of highly similar tandem domain repeats might not be as widespread as previously suggested, or that mechanisms for misfolding and aggregation resistance might have evolved in certain domain families.

I have further observed a widespread amino acid composition bias in highly similar tandem domain repeats across many different Pfam families, with striking correlations between the bias and the sequence identities of adjacent domains. The molecular origin and consequences of this sequence bias for protein domains is not yet clearly understood. I explored several properties of amino acid side-chains, such as the side-chain entropy, but they did not appear to generally correlate with the type of amino acid bias observed, and a direct link between protein aggregation-resistance could not be found. In the next chapter (Chapter 3), I explore further this sequence composition bias in a subset of bacterial surface proteins.

In this study, I used two different tandem domain repeat detection approaches to compensate for their limitations. *De novo* tandem repeat detection methods use heuristics-based algorithms to avoid prohibitive running time complexities, which often cause unpredictable instabilities in the number of repeats and their boundaries in the multiple sequence alignment and reduces their scope to less frequent highly similar tandem domain repeats (usually on the order of 90% sequence identity). In addition, large domain-size tandem repeats do not always correspond to globular domains — the main focus of this study — and repetitive regions can also be disordered or fold into other repetitive proteins such as Ankyrin and Leucine-rich repeats. The repeat size used in this study (50-200 residues) captures average length domains, around 100 residues, but also allows the detection of relevant double domain repeats (repeating sequences of two domains), for example in the SasG protein.

The detection approach based on Pfam aimed to tackle some of these limitations by using pre-existing domain family definitions. This approach is more robust in the number and alignment of repeats detected and it extends the search to more remote domain repeat homologs (below 30% sequence identity), but it is limited to the identification of previously known domains. Although the Pfam

coverage of tandem domain repeat regions has increased in the course of this study, it remains 20% lower than the Pfam average and many domain repeats remain uncharacterised. One of the reasons for this coverage disparity is the biased sequences of tandem domain repeats, which cause disorder prediction methods like IUPred to wrongly predict them as disordered regions, and complicates their alignment by sequence-based tools like HMMER.

Even though tandem domain repeats are rare in proteins, they participate in a wide range of important biological functions and test our assumptions of protein folding and sequence evolution. This comprehensive study improves our knowledge of the types of natural tandem domain repeats, their roles in proteins, and their distribution across domain families and organisms; and serves as the basis for further investigations in the following chapters of this thesis.

2.5 Methods

2.5.1 Tandem repeat detection with T-REKS

The T-REKS tool (Jorda and Kajava, 2009) is available as a web-server and Java application, consisting of a graphical user interface (GUI) for the analysis of small subsets of sequences and a command line interface (CLI) for large-scale analyses of millions of proteins. The tool accepts sequences in a FASTA file as input and outputs a table of repeats with a row for each sequence and a multiple sequence alignment of the repeats; and it takes several parameters as options, such as the minimum percentage of identity between repeats and the maximum percentage of gaps allowed. To find domain-size repeats of length 50–200 residues, the output table of repeats was filtered with a BASH script after each T-REKS run. T-REKS is very fast: the running time to process half a million protein sequences in the SwissProt database is in the order of 5 minutes in a laptop with a common 4-core 2.9 GHz Intel processor.

During initial calculations with T-REKS across large datasets of proteins, I noticed an unexpected failure mode in certain sequences that caused the entire calculation to stop. As an instrumental tool to this project, I contacted Dr. Andrey Kajava at the Montpellier Cell Biology Research Center (CRBM) reporting the issue. They helped me understand the error (caused by their custom sequence aligner implementation), suggested that I use an external sequence alignment tool

(MUSCLE) to fix it, and sent me a copy of the T-REKS source code. I modified the T-REKS source code to include the suggested changes and bug fixes and improved its command line interface (input and output options), error reporting and efficiency for large-scale calculations. I used this modified version of T-REKS for the results presented here, and released the modified source code in a GitHub repository: https://github.com/lafita/treks-hpc.

Database versions

I originally performed the survey of tandem domain repeats with T-REKS at the start of the project (May 2017) using UniProt version 2017_05 (88 million sequences), and I found repeats in 50 thousand proteins with a sequence identity threshold of 80%. I later repeated the survey in July 2019 using the new custom version of T-REKS, a higher sequence identity threshold of 90%, and the newer UniProt version 2019_06 with double the number of sequences (160 million). The results from the 2019 survey are presented in this chapter.

For the structural coverage of tandem domain repeats (Figure 2.3), I used the original survey results in UniProt 2017_05 and the PDB (and associated SIFTS dataset) released in May 2017 (2017/05/07), containing around 40 thousand unique protein chains. For the Pfam coverage of T-REKS tandem domain-size repeats (Figure 2.2), I originally used Pfam version 31.0 (released in March 2017), and later updated the coverage with Pfam version 33.1 (released in May 2020) to analyse the coverage improvement over the course of the project.

I used Pfam version 31.0 (March 2017) and its database of sequences (based on UniProt Reference Proteomes) for the Pfam-based tandem domain repeat detection (Figures 2.4 and 2.5); the new families created in this study were therefore not part of this analysis. I only used Pfam families of type "Domain" and ignored the others (a total of 5,130 out of 17 thousand Pfam families).

2.5.2 Building Pfam families from tandem repeats

Tandem repeats identified by T-REKS and without Pfam annotations were first clustered using BLAST (Altschul *et al.*, 1997) to identify groups of similar sequences, and representative repeat sequences from each group were selected as potential candidates for building new families.

The process of creating a new family starts from the multiple sequence alignment of repeats generated by T-REKS, used as the initial SEED alignment of the family. Pfam internal tools are then used to build an initial HMM model and homologous domains are searched using the HMMER tool across protein sequences in Pfam (based on the UniProt Reference Proteomes). The multiple sequence alignment of homologs generated by HMMER is then manually checked, removing partial and redundant sequences, and saved as the new SEED alignment of the family. Multiple iterations of this procedure might be needed to find more remote homologs of the family. If the family has a sufficient number of members (above one hundred and commonly on the order of thousands) and it passes the Pfam internal quality checks, the family is added to Pfam with a new identifier.

There are several challenges specific to building families from multiple sequence alignments of tandem repeats. The domain boundaries of the repetitive region detected by T-REKS (and other tandem repeat detection methods) are often out of phase compared to the real structural boundaries of domains. To correct domain boundaries, overlapping regions to homologous families with known structural boundaries (found using Pfam domain hit overlaps) are used.

2.5.3 Calculation of domain sequence properties

Sequence identity

The percentage of sequence identity between two domains is calculated using their sequence alignment, obtained from T-REKS repeat alignments and Pfam family alignments generated by HMMER, and ignoring gap positions. Identical amino acids in a position are counted and divided by the number of non-gapped positions in the alignment.

Tandem domains were classified into low identity and high identity using their sequence identity to immediately adjacent domains. Domains at the termini of tandem regions only have one adjacent domain, but domains at the middle of tandem regions have two adjacent domains, so the highest percentage of identity was assigned.

Sequence bias

The sequence bias metric used is the absolute Shannon entropy of the sequence. The Shannon entropy H of a protein sequence S is computed from the individual frequencies f_i of the 20 amino acids as following:

$$H(S) = -\sum_{i=1}^{20} f_i \log_2(f_i) \tag{2.1}$$

The Shannon entropy of a sequence is in the range $[0, \log_2(\frac{1}{20})]$. Values close to 0 represent highly skewed sequence compositions towards one or few amino acids; values close to $\log_2(\frac{1}{20})$ or 4.32 represent a uniform distribution of amino acids close to 5% (1/20) each.

Side-chain entropy

The side-chain entropy S_{sc} of a sequence S is estimated by multiplying the individual amino acid frequencies f_i with the side-chain entropy of amino acid side-chains δ_i from Table 2.5 as following:

$$S_{sc}(S) = -\sum_{i=1}^{20} f_i \delta_i \tag{2.2}$$

Domain length

The length of domains is extracted from multiple sequence alignments produced by the HMMER tool using Pfam family models. Since alignments tend to be unreliable at the domain boundaries, a correction to the aligned length was applied in order to avoid artifacts by adding terminal HMM model gaps to the total domain length.

For every HMM position, the gap fraction was calculated as the number of gaps in each HMM position divided by the total number of sequences. The average number of insertions was calculated by summing up insertion lengths between every HMM position and dividing by the total number of sequences.

TABLE 2.5 **Table of amino acid side-chain entropies estimated by Pickett and Sternberg (1993)**

Amino acid	Entropy (kcal/mol)
A	0.00
R	-2.03
N	-1.57
D	-1.25
C	-0.55
Q	-2.11
E	-1.81
G	0.00
H	-0.96
I	-0.89
L	-0.78
K	-1.94
M	-1.61
F	-0.58
P	0.00
S	-1.71
T	-1.63
W	-0.97
Y	-0.98
V	-0.51

Chapter 3

Discovery of bacterial Periscope proteins

> *"Simply ignoring repeats is not an option, as this creates problems of its own and may mean that important biological phenomena are missed."*
>
> - Treangen and Salzberg (2012): "Repetitive DNA and next-generation sequencing: Computational challenges and solutions"

In this chapter, I describe the study of Periscope proteins, a project in collaboration with the lab of Prof. Jennifer Potts at the University of York. I computationally discover new Periscope proteins — a new class of repetitive bacterial surface proteins — from a dataset of high-quality bacterial genomes, and characterise their sequence composition biases and repeat number variability at the genomic level.

Some sections and figures in this chapter are included in a recent manuscript available in *bioRxiv* and recently submitted to a scientific journal: Whelan *et al.* (2020). I am co-first author on the paper and actively participated in manuscript writing and figure generation.

3.1 Introduction

Bacteria are found in the most inhospitable environments — from extreme temperature and salinity conditions to hostile habitats such as our gastrointestinal system — and often infect humans and other organisms as they face the perpetual struggle for survival. As a result, and despite their apparent simplicity relative to eukaryotic organisms, bacteria have evolved sophisticated mechanisms to rapidly respond to sudden changes in their surroundings and adapt to new environments. There are many widespread bacterial adaptation mechanisms reported and well characterised in the literature, such as "phase variation" (Phillips *et al.*, 2019).

A less well-studied mechanism is length variation in bacterial surface proteins. Several bacterial surface proteins implicated in host colonisation and biofilm formation, such as the surface proteins Rib and SasG (Figure 3.1), contain highly similar tandem domain repeats. The length variability in these surface proteins could confer a mechanism to adapt to new selection pressures, for example to evade the host immune system and to regulate surface interactions (Gravekamp *et al.*, 1996).

Several studies previously reported length variability in bacterial surface proteins, corresponding to different numbers of tandem domain repeats. Wästfelt *et al.* (1996) observed repeat number variability in the *S. agalactiae* surface protein Rib using Polymerase Chain Reaction (PCR) and Western blot techniques. They attributed the different bands observed in the protein Western blot to hydrolysis of an acid-labile Asp-Pro bond in the repeats, but for the results of the PCR they noted that: "An interesting observation made during the PCR analysis was that the PCR product not only contained the main band but also gave rise to a ladder of bands with a size difference of 237bp, corresponding to one repeat. This ladder could be the result of slippage of Taq polymerase during replication, due to the unique repetitive structure of the rib gene."

Later studies by Lachenauer *et al.* (2000) and Roche *et al.* (2003) confirmed the repeat number variability observations in the surface protein Rib and demonstrated that the mechanism is widespread across other repetitive bacterial surface proteins, such as *S. aureus* surface protein SasG. Even though polymerase slippage and homologous recombination were suggested as likely mechanisms for this repeat number variations, none of the studies found conclusive evidence due to experimental limitations of PCR and Western blot techniques.

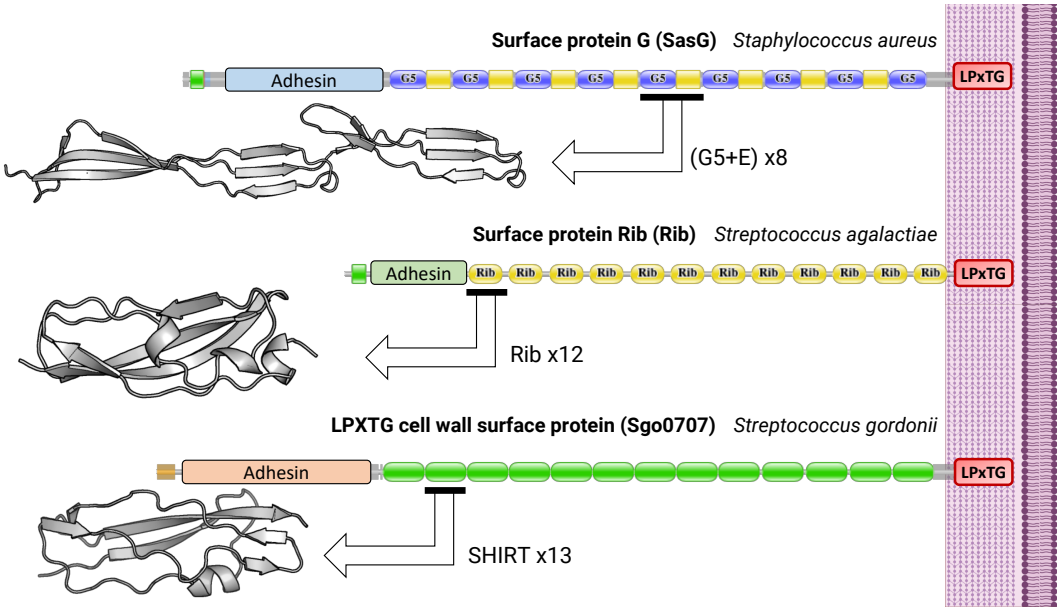

FIGURE 3.1 Three examples of repetitive bacterial surface proteins with highly similar tandem domain repeats. Proteins SasG (UniProt:Q2G2B2), Rib (UniProt:P72362) and Sgo_0707 (UniProt:A8AW49) are anchored at the bacterial cell wall of *S. aureus*, *S. agalactiae* and *S. gordonii*, respectively. The proteins are represented as domain architecture diagrams, each box corresponding to one protein domain colored and labelled by its Pfam family. The location of the terminal adhesin domains and LPxTG cell-wall anchoring motif are also shown, but do not correspond to exact domain locations. Structures of the repeating domain units for each protein are shown in cartoon representation: G5+E (PDB:3TIP), Rib (PDB:6S5X) and SHIRT (PDB:7AVK).

Years later, experimental structures by the lab of Prof. Jennifer Potts at the University of York showed that these repeating elements in bacterial surface proteins fold into stable globular domains (Gruszka *et al.*, 2012; Whelan *et al.*, 2019; Whelan *et al.*, 2020). This observation shed light into their role forming elongated and rigid stalks that project the protein out of the bacterial cell surface (forming thin fibrils). It also became apparent how their repeat number variation modulates their surface exposure and therefore their role in host invasion and biofilm formation.

In close collaboration with the group of Prof. Jennifer Potts, we generalised these ideas into a new class of proteins that we named "Periscope proteins" for their ability to differentially surface out of the bacteria and interact with their environment. In this chapter, I computationally discover new members of the Periscope proteins class from bacterial genomes and characterise their domain repeat number variability.

In the following introductory subsections, I review some important properties of bacterial surface proteins and genomic mechanisms of bacterial adaptation. I also define Periscope proteins and their characteristic domain architecture, and describe the dataset of bacterial genomes that I used to computationally discover them.

3.1.1 Bacterial surface proteins

As the primary interface between bacterial cells and their environment, surface proteins are critical to bacterial survival: they participate in important functions, such as cell adherence, host invasion, biofilm formation and signalling, and are subject to strong evolutionary pressures (Navarre and Schneewind, 1999; Foster *et al.*, 2014). They also interact with the host immune system and other proteins and molecules in the environment.

Bacteria are classified into Gram-positive and Gram-negative types in relation to the composition of their cellular envelope, which drastically impacts the types of surface proteins they present. Gram-positive bacteria are surrounded by an inner membrane and a thick peptidoglycan cell wall layer, which is in direct contact with the extracellular environment. Gram-negative bacteria are also surrounded by an inner membrane, but have an additional outer membrane and only a thin peptidoglycan layer in between. Their name arises from the gram-

staining method of bacterial differentiation, where Gram-positive bacteria retain the crystal violet stain and Gram-negative bacteria do not.

Surface proteins bind to the bacterial surface through various mechanisms: they can be attached to the bacterial membrane through transmembrane regions or lipoproteins, or, in Gram-positive bacteria, they can be anchored to the cell-wall through covalent and noncovalent interactions to its associated proteins (Desvaux *et al.*, 2006). Two common cell-wall anchoring mechanisms include Sortase-mediated covalent bonding to peptidoglycan through C-terminal LPxTG motifs (or related) and integration into the S-layer (an additional protein coating present in some bacteria) through terminal S-layer homology (SLH) domains. Computational methods have been developed to predict protein surface localisation based on these cell surface attachment motifs and domains in proteins, and include tools such as PSORTb (Gardy *et al.*, 2003) and Inmembrane (Perry and Ho, 2013).

Bacterial surface proteins show a complex and wide diversity of domain architectures and sequences, including ubiquitous long and short repetitions and low complexity segments such as Serine-rich regions (Fischetti, 2019). Several bacterial surface proteins with long arrays of tandem domain repeats have been described in the literature. Three examples are shown in Figure 3.1: the surface protein G (SasG) containing a double domain repeat, and the surface proteins Rib and Sgo0707 with single domain repeats. Domain repeat regions in these proteins also show highly biased amino acid sequences (Gruszka *et al.*, 2012), and popular disorder prediction tools like IUPred (Dosztányi *et al.*, 2005) confidently mispredict them as disordered even though they fold into stable globular domains.

3.1.2 Bacterial adaptation and phase variation

Bacteria have evolved various strategies to rapidly respond and adapt to changing environments and adverse conditions. In addition to classical control by regulation of gene expression, bacteria exploit mechanisms that give rise to random variation to facilitate adaptation, such as phase and antigenic variation (Phillips *et al.*, 2019).

Phase variation consists in switching ON and OFF sets of genes through genomic mutation mechanisms such as DNA inversions (J. Li *et al.*, 2016), homologous recombination (Vink *et al.*, 2012), and replication slippage (Zhou *et al.*, 2014).

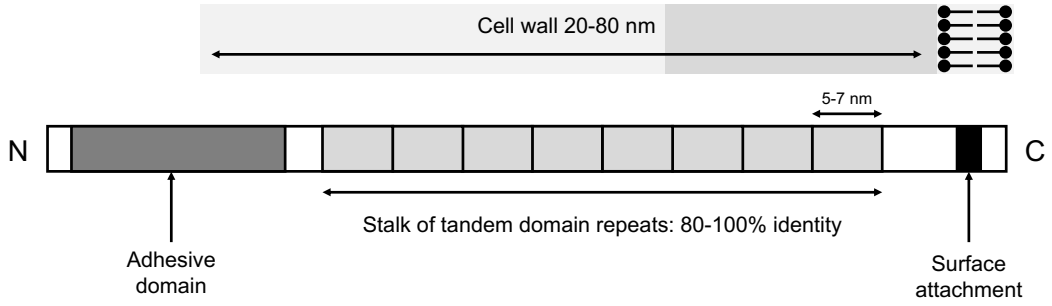

FIGURE 3.2 Schematic representation of the Periscope protein architecture. The size of a typical bacterial cell envelope is shown on top for reference. Varying numbers of stalk domains result in differential exposure of the protein out of the bacterial cell envelope.

Many of these mechanisms are mediated through short sequence repeats (SSRs) flanking specific regions of bacterial genomes, which promote spontaneous duplications, deletions and repeat number expansions and contractions (Moxon *et al.*, 2006).

The number of tandem domain repeats in bacterial surface proteins has been associated to phenotypic variability in bacterial strains. For example, in the surface protein Rib, Gravekamp *et al.* (1996) found that repeat variation is linked to antigenicity and protective epitopes, and Almeida *et al.* (2017) investigated bacterial pathogenicity using large-scale genomic sequencing datasets and found that disease-causing Group B Streptococcus (GBS) strains had a significantly lower number of repeats in the Rib protein.

3.1.3 Periscope proteins: definition and architecture

Together with Jennifer Potts, we defined a new class of proteins, named "Periscope proteins", that are located at the surface of bacterial cells and have the following domain architecture and function properties: i) they are attached at the cell surface through membrane or cell-wall anchoring mechanisms; ii) contain a central stalk of nearly identical domain repeats that are variable in number and that project straight out of the bacterial surface; and iii) they interact with the bacterial surroundings through adhesive or other effector domains, located at the protein terminal end farther from the bacterial surface (Figure 3.2).

The name "Periscope" is intended to allude to their functional and structural

similarity to their namesake instruments in submarines. Large numbers of tandem domain repeats in Periscope proteins enable them to surface out of the bacterial envelope and interact with the environment. At the same time, their repetitive sequences provide a molecular mechanism to effectively generate phenotypic variability, modulating the protein length and therefore surface exposure by changing the number of repeats in Periscope genes. By folding as globular units, tandem domain repeats in Periscope stalks form a rigid linear rod-like structure, the length of which varies directly proportional to the total number of tandem repeats (Gruszka *et al.*, 2012; Whelan *et al.*, 2019).

Only a handful of proteins with the Periscope domain architecture and function were known previous to this work, and their similarity, evolution and domain repeat variability had not been thoroughly studied. Here, I identify new members of the Periscope proteins class from bacterial proteins and study their properties.

3.1.4 The NCTC3000 dataset of bacterial genomes

The large amount of sequenced bacterial genomes available nowadays permit extremely detailed analyses of genomic and proteomic variations. The relative simplicity of bacterial genomes, in particular the absence of gene splicing mechanisms, is also an advantage to interpret genomic variants. However, repetitive genomic regions, specially those with nearly identical repeats, pose major challenges for genomic studies; short-read sequencing technologies like Illumina, the most commonly used, have fundamental limitations to correctly assemble repetitive genomic regions (Treangen and Salzberg, 2012).

Tandem domain repeats in Periscope proteins are both very long and highly similar (around 300 bases and adjacent repeats in the range 95-100% sequence identity) and are therefore susceptible to high assembly errors by short-read technologies like Illumina; with an average read length of 100 bases, these short reads do not cover the entire repetitive region (nor even a single individual repeat) and therefore the repeat number and order cannot be accurately assigned. Almeida *et al.* (2017) managed to get around this problem by estimating the number of repeats in the gene using the normalised read coverage of repetitive genes, similar to techniques used to estimate gene copy number variations (CNV). This is, however, an indirect and noisy measure of the number of repeats, with many con-

founding factors, such as the sequencing depth and number of repeats in the gene, and it does not permit sensitive individual variability analyses between bacterial strains. To avoid these problems, I decided to use bacterial genomes sequenced with long-read technologies, which are several kB long and span whole Periscope genes; these allow to reliably assemble repetitive genomic regions and produce accurate numbers of repeats in assembled genes and proteins (Tørresen *et al.*, 2019).

I chose a specialised dataset of bacterial strains sequenced using the PacBio sequencing technology to carry out the identification and analysis of Periscope proteins. The NCTC3000 dataset is an ongoing project led by the Sanger Institute providing high quality annotated genome assemblies for 3,000 bacterial strains from Public Health England's National Collection of Type Cultures (NCTC).[1] The NCTC3000 dataset contains, however, only a limited number of bacterial species (Figure 3.3) and the project is still ongoing, so only a fraction of the genomes have been sequenced and annotated (734 as of October 2019). Its advantages include high quality assembled proteins from well-known and studied bacterial strains and a direct link between sequencing reads, assembled genes and proteins, which is not always available in proteins from other databases like UniProt.

3.2 Identification of Periscope proteins

One of the characteristic properties of Periscope proteins is the presence of nearly identical tandem domain repeats, so-called stalk domains. In order to identify new Periscope proteins across the NCTC3000 bacterial genomes, I focused on searching for domain-size highly similar repeats, presumably corresponding to stalk domains.

3.2.1 Detection of stalk domain repeats

First, I extracted a dataset of 2.5 million proteins from 734 genomes in the NCTC3000 dataset (details in the Methods section). I then ran the T-REKS tool (Jorda and Kajava, 2009) across all proteins in the dataset using a sequence

[1] NCTC3000: https://www.sanger.ac.uk/resources/downloads/bacteria/nctc

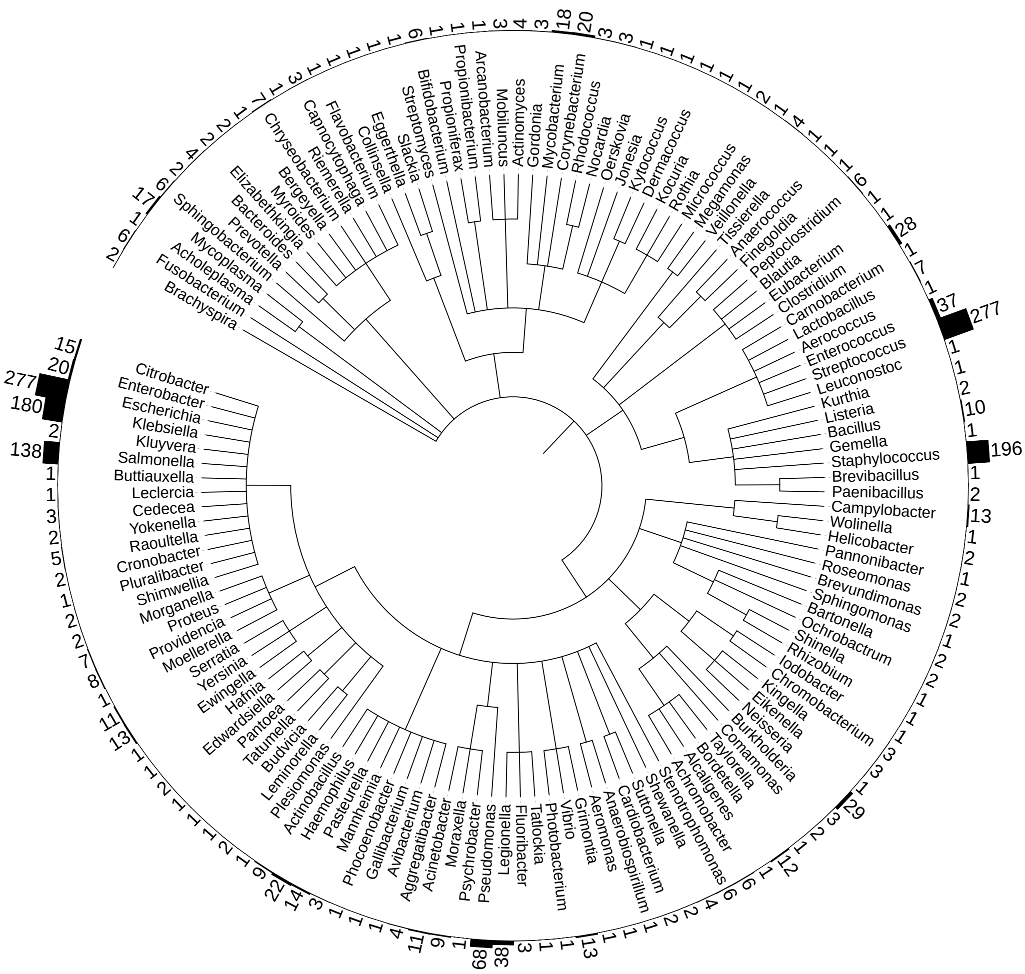

FIGURE 3.3 Taxonomic tree of species in the NCTC3000 genomes. Tree shown at the genus level, with bars and numbers indicating the number of genomes (strains) for each genus. Taxonomic tree was downloaded from the NCBI Taxonomy database (Federhen, 2012) and visualised using iTOL (Letunic and Bork, 2019).

identity threshold of 80% [2] and a repeat length between 50-200 residues.

I found a total of 1,576 proteins with repeats (0.06% of all proteins in the dataset), but these repetitive proteins were widespread across genomes: 538 strains contain at least one protein with repeats and 125 strains contain 5 or more. I further clustered the sequences of repeats (details in the Methods section) and obtained a total of 124 unique groups (Figure 3.4). Most clusters are homologous to existing Pfam families; the other largest clusters without Pfam annotations were further iterated to find more sequences in UniProt and, if large enough, built as new families. Pfam domain annotations reveal that these repeats mostly fold into globular domains, as in previously known Periscope proteins, and are part of a wide range of secondary structure and fold types. Most of the domain families I find correspond to Immunoglobulin-like β-sandwich folds in the E-set clan (Pfam: CL0159), but I also find domain families with mirror β-grasp folds in the new MBG clan (Pfam: CL0682), and left-handed three-helix bundles in the bacterial immunoglobulin/albumin-binding clan (Pfam:CL0598).

3.2.2 Putative Periscope protein groups

I further clustered full-length repetitive proteins into unique groups of nearly identical sequences, but variable lengths (details in the Methods section). I identified a total of 180 unique groups, 84 of which exhibit repeat number variations according to the T-REKS repeat detection results. I manually inspected the protein groups that exhibit repeat number variability, looking for bacterial surface-associated domains and motifs, and the typical Periscope protein architecture. I confirmed that at least 56 of these protein groups are examples of Periscope proteins, 30 of which could be further assigned to recognisable gene names (Table 3.1).

In Table 3.1, I find previously known Periscope proteins: surface proteins SasG, SasY, Rib and Sgo_0707; as well as other putative Periscope proteins well-characterised in the literature, for example the surface agglutinin CdrA from *Pseudomonas aeruginosa* (Borlee *et al.*, 2010), the CshA adhesin from *S. gordonii* (McNab *et al.*, 1999; Back *et al.*, 2017), or the PavB adhesin from *S. pneumoniae* (Jensch *et al.*, 2010). The majority of the manually confirmed Periscope protein

[2]The sequence identity threshold used here is lower than the 90% used in previous analyses due to the lower number of proteins in the NCTC3000 dataset.

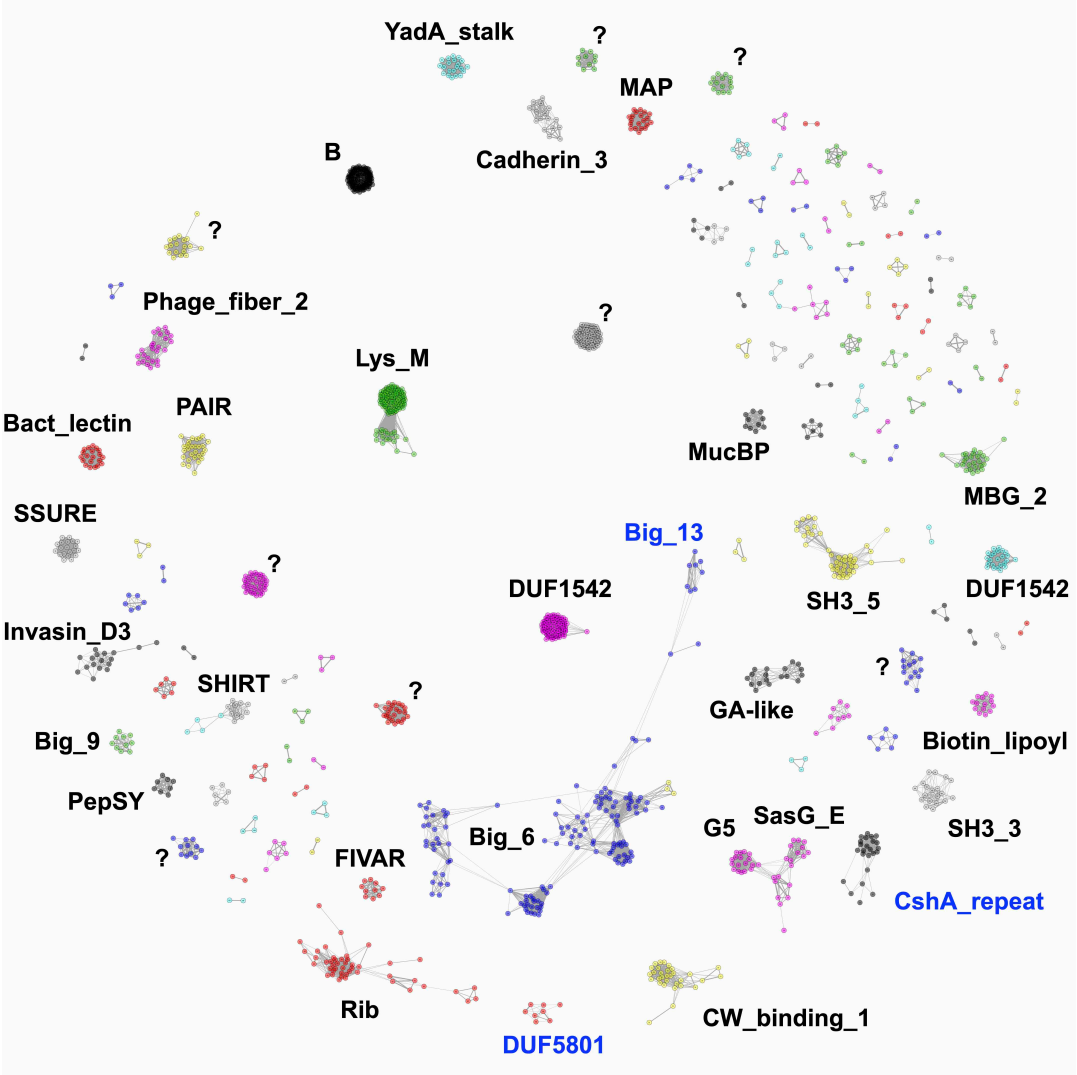

FIGURE 3.4 Sequence clustering of domain-size tandem highly similar repeats identified across proteins of the NCTC3000 genomes. Largest clusters annotated with Pfam or marked with "?" if unknown. Names in blue represent new Pfam families built from these sequence clusters. Network visualised using the `igraph` package in `R`. Figure adapted from Whelan *et al.* (2020).

TABLE 3.1 Putative Periscope proteins identified in the NCTC3000 strains. Columns show the total number of proteins in the cluster (size), the minimum and maximum number of repeats among proteins in the cluster (minrep, maxrep), the repeat length (replen), average T-REKS percentage of identity (psim), and Pfam annotation of the repeat sequence (pfam_rep). Table sorted by Name, with unknowns (UNK) at the bottom.

Name	UniProt ID	size	minrep	maxrep	replen	psim	pfam.rep
Bag	A0A4V0C2I4	2	11	12	89	1.00	CFSR
BapA	A0A5E9KBQ0	23	2	15	83	0.89	Big_6
Bca_2	A0A4U9ZPF1	4	9	24	134	0.98	Rib
CdrA	A0A485CMM4	24	3	15	81	0.88	MBG_2
CshA	A0A0F2D244	10	3	14	101	0.84	CshA_repeat
Emm23	Q9RHV2	3	2	3	63	0.86	
Epf	Q48UF3	10	3	10	81	0.89	DUF1542
Esp	Q9Z4N7	12	3	10	82	0.87	Rib
Fap	A0A1Q1FVK8	5	2	5	131	0.93	DUF11
HxuA	A0A379IT80	2	9	10	88	0.82	YDG
LytD	A0A2X2YDI6	4	3	6	85	0.87	SH3_3
PavB	C1C9Y2	5	2	9	152	0.91	SSURE
PavB_2	A0A3R9HDH8	8	2	6	150	0.89	SSURE
PavB_3	A0A1X1K6M7	2	2	4	149	0.95	SSURE
R28	Q48S64	7	2	15	79	0.96	Rib
Rib	A0A1A9DZG2	5	2	7	82	0.92	Rib
Rib_2	A0A4V0EE13	2	7	12	76	0.92	Rib
SasG	Q2G2B2	34	3	13	128	0.84	G5, SasG_E
SasG_2	T1YDP9	5	3	13	128	0.94	G5, SasG_E
SasY	A0A4V0BAN3	6	3	15	87	0.91	Big_6
Sgo0707	A0A3R9JLL4	7	7	13	84	0.91	SHIRT
ShdA	A0A3U7M2R1	15	4	6	59	0.85	PATR
ShdA_2	A0A400S776	8	2	14	64	0.89	PATR
Spa	M4MB15	97	2	7	59	0.87	B
Spg	A0A4V0F6C4	10	2	4	75	0.85	GA-like
SraP	A0A376H473	14	2	46	98	0.92	
SraP_2	A0A380G3G2	3	2	14	89	0.94	He_PIG
Vwbl	Q6QQB0	2	7	9	67	0.87	SSSPR-51
YeeJ	A0A192CME6	3	2	3	119	0.94	Invasin_D3
ZmpC	F0I817	10	2	10	96	0.89	G5
UNK	A0A507HZU9	9	3	4	154	0.91	
UNK	A0A381IWY6	6	3	9	96	0.94	
UNK_Big13	Q0B1Q8	17	2	31	103	0.93	Big_13
UNK_Big13	A0A2I5HP40	6	2	9	103	0.89	Big_13
UNK_Big13	A0A4P9IZ13	3	9	16	95	0.95	Big_13
UNK_Big3	E2Z005	5	3	6	73	0.83	Big_3
UNK_Big6	A0A2T4PXZ6	14	2	21	87	0.92	Big_6
UNK_Big6	A0A0M2AG42	4	4	5	90	0.88	Big_6
UNK_DUF1542		87	2	4	78	0.95	DUF1542
UNK_DUF1542	A0A5S4TRV8	3	5	7	81	0.87	DUF1542
UNK_DUF1542	A0A4Q9WP12	2	5	6	77	0.91	DUF1542
UNK_G5	A0A4U9ZZC2	6	3	5	79	0.90	G5
UNK_Invasin	A0A376HGL9	5	3	12	97	0.84	Big_1
UNK_Invasin	A0A377P116	2	3	6	101	0.90	Invasin_D3
UNK_Invasin	A0A380CN38	2	7	8	103	0.81	Invasin_D3
UNK_Lyase	A0A449G498	8	3	7	70	0.89	Lyase_8_C
UNK_Lyase	A0A4U9PE62	3	5	6	69	0.86	FIVAR
UNK_MucBP	A0A5S4TIX9	5	4	7	81	0.85	MucBP
UNK_MucBP	A0A0E2I0J3	3	2	3	75	0.83	MucBP
UNK_Rib	A0A2T4QWP0	3	5	12	86	0.92	Rib
UNK_Rib	H2A912	3	5	14	79	0.91	Rib
UNK_SHIRT	A0A1X1FZN4	2	4	7	84	0.84	Cna_B
UNK_SpaA	A0A455TW44	7	2	4	93	0.89	SpaA
UNK_SpaA	A0A2J9ETF2	7	3	17	112	0.87	SpaA
UNK_SpaA	A0A3R9LMV6	2	7	8	114	0.90	SpaA
UNK_YadA	A0A3T3ETZ9	12	3	4	90	0.86	YadA_stalk

groups (40/56) have at least one Gene Ontology (GO) term for surface localisation associated with the cell-wall (GO:0005618) or bacterial membrane (GO:0016020, GO:0016021); and several are annotated as having cell adhesion (GO:0007155) function.

3.2.3 Sequence bias in Periscope genes

Among the group of potential Periscope proteins in Table 3.1, I identified a strong sequence bias in the protein sequences and coding genes. Initially, I observed a skewed composition of Adenine over Thymine, meaning that the fraction of A was higher than that of T in the genes, even though the proportion of A and T is expected to be roughly the same, as observed in random genes from the same genomes (Figure B.1). Periscope genes appear to always have positive and extreme AT skews, reaching values of 50% in some genes, while random genes have a wide spectrum of positive and negative values and very rarely above 30%.

Upon closer inspection, the AT skew does not seem to be determined by the wobble base in codons, but rather it is strongest in the first and second codon positions (Figure B.2), suggesting that the AT skew is caused by a specific bias in the amino acid composition and not by a saturation of Adenine in the genes. The amino acid composition in Periscope genes, however, does not show a single or a small group of enriched amino acids, but rather a wide spectrum of amino acids that appear enriched or depleted in different Periscope genes, but not in others, for example Ala, Gly, Asn, Pro and Thr (Figure 3.5).

Interestingly, when looking at the profile of the amino acid bias along the sequences of Periscope proteins, I observed that the bias is mainly found at the repetitive region of the proteins, where the stalk domain repeats are found (Figure 3.6). The amino acid bias in the Rib, SasG and SraP proteins is very strong, with few amino acids enriched and depleted in the repetitive region. However, the types of amino acids that are enriched and depleted are different in the three proteins. In Rib domains, D, P, T and V are highly enriched, while A, I, L and S are depleted (Figure 3.7). In SasG, the enriched amino acids are E, P, T, and K; and N, A, L and S are depleted (Figure B.5). In SraP, however, A is the most enriched amino acid (which is depleted in CdrA and Rib). In CdrA, a Gram-negative Periscope protein, the enriched amino acids are G, N, and L; while the depleted amino acids include P and T (Figure B.6).

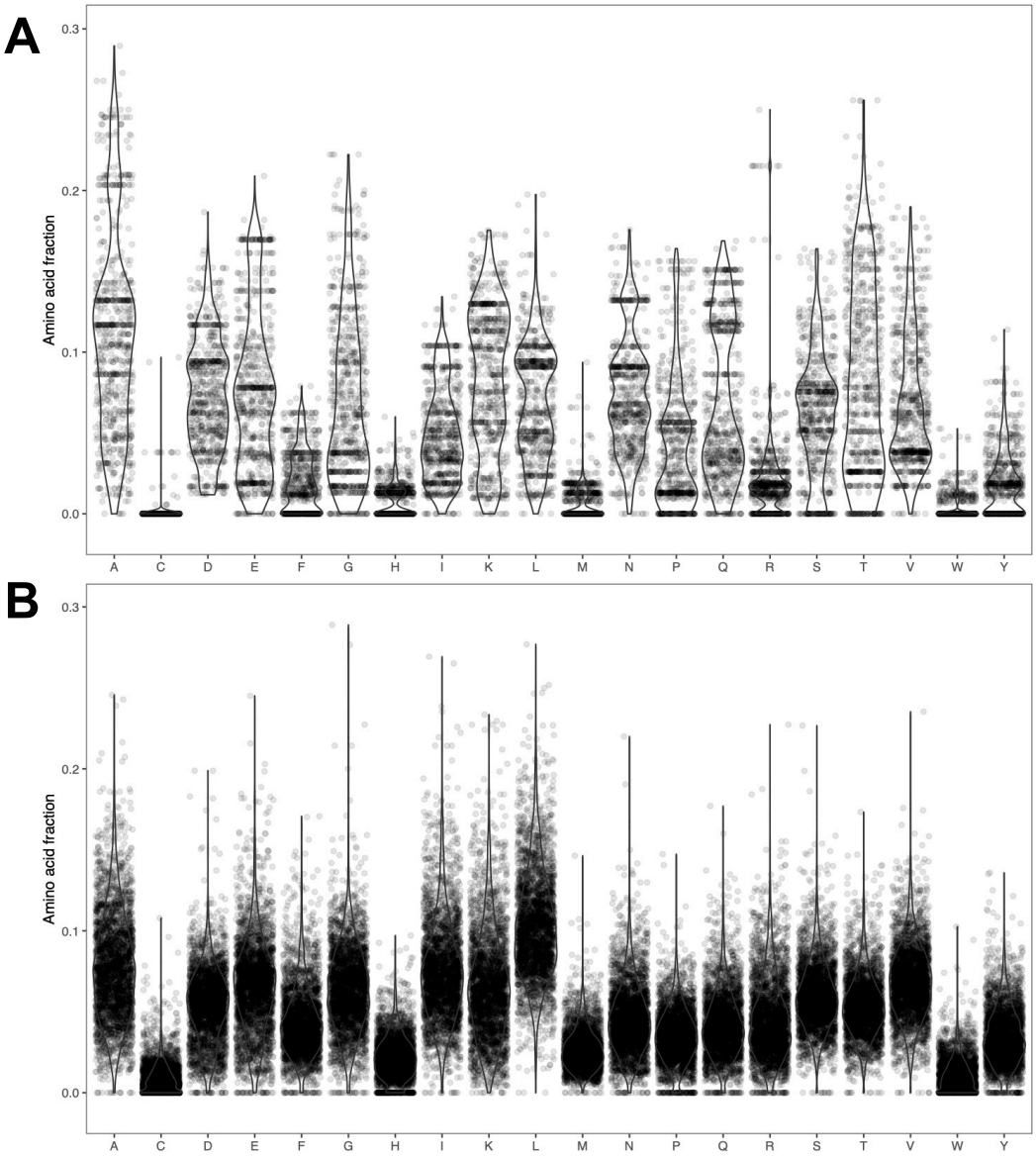

FIGURE 3.5 Amino acid composition of stalk domain repeats in Periscope proteins. Violin plots of amino acid composition of the sequences of stalk domain repeats in Periscope proteins (A) and a subset of random proteins from the same genomes as the background (B).

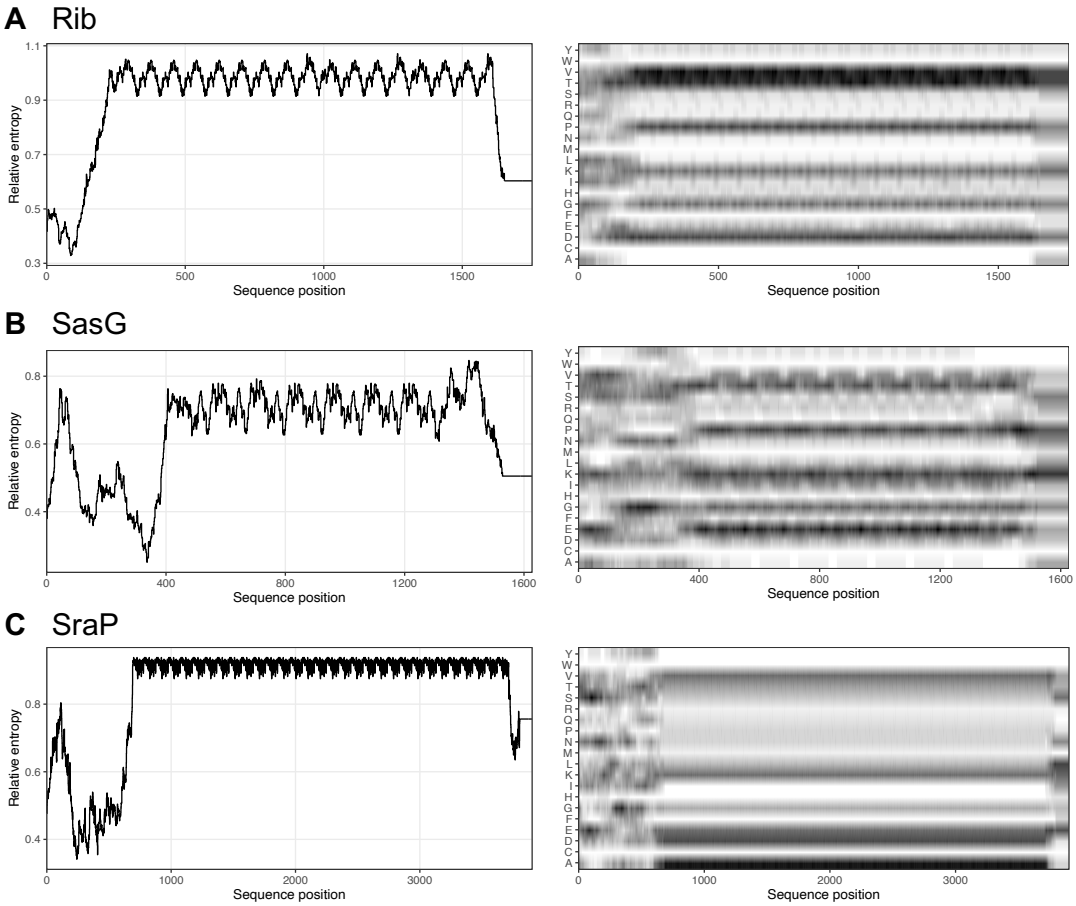

FIGURE 3.6 Sequence bias and amino acid profiles of Periscope proteins. Sequence bias along the protein measured as the relative entropy (left) and amino acid composition measured as the fraction of each amino acid (right), for the Periscope proteins Rib (A), SasG (B) and SraP (C). Amino acid fractions are colored in a grey scale (white to black) in the 0-1 value range. Sequence bias and amino acid fractions are calculated using a rolling average with a window size of 100 residues. Higher entropy values represent more biased compositions.

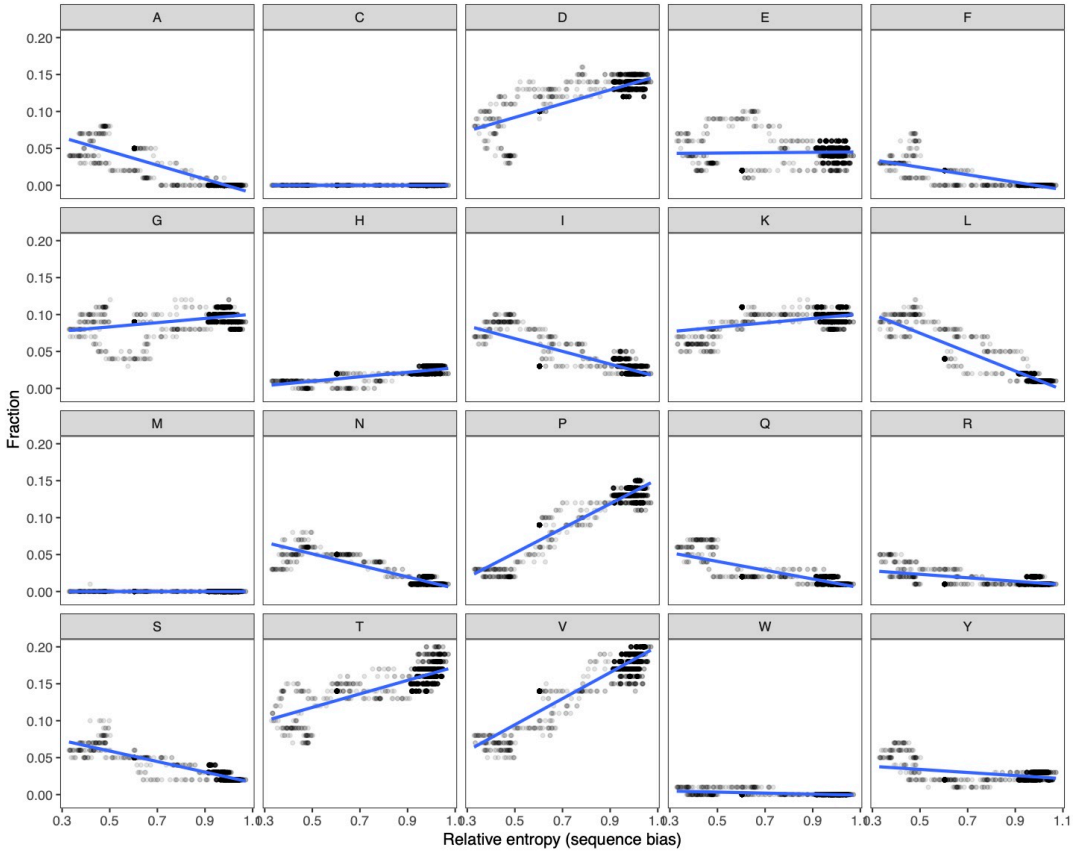

FIGURE 3.7 Sequence bias and amino acid correlations in the surface protein Rib. Analysis of the correlation between the protein sequence bias (x-axis) and the amino acid fraction (y-axis), split for each of the 20 amino acids. Linear fit for each plot is shown as a blue line.

3.3 Variability of stalk domain repeats

Previous studies reported length variation through domain repeat number changes in Periscope proteins, and hypothesised that this change occurred at the DNA level, possibly through DNA recombination or a similar molecular mechanism. My goal now was to characterise the repeat variation in Periscope proteins at the genomic level, in order to measure the frequency and rate of repeat number changes.

3.3.1 Within strain variability

The first approach I explored was the use of raw sequencing reads to detect repeat number variations within the same genome. I extracted individual PacBio long raw reads from each bacterial genome that covered the entire Periscope gene, in order to count the number of repeats directly in the raw read without the need to assemble them. This type of analysis would be the *in silico* equivalent of an experimental Southern blot, and was used to detect repeat number variation of Periscope genes within the same bacterial strain or colony.

I selected four Periscope genes (SasG, SasY, Rib and Sgo_0707) to study their repeat number variation in raw PacBio reads from a total of 82 genomic strains. I extracted raw reads covering the entire Periscope gene of interest from each genome and counted the number of repeats using three methods: the gene length, the repetitive region length, and the number of hits to the repeat sequence (details in the Methods section). For each genomic strain, I created a plot of repeat number estimates like the one shown in Figure 3.8 and manually inspected if the three repeat number counting strategies differed in any of the raw reads. I also manually compared reads with an estimated number of repeats that differed to the assembled gene using dotplots to confirm if the estimated repeat number change was correct, but none of the examples found turned out to be a true example of repeat number changes and they likely corresponded to sequencing artefacts.

The analysis of repeat number variation within the same genome turned out to be challenging. Sequencing errors, including indels and nucleotide assignment errors, caused by limitations of the sequencing technology, are common in single molecule sequencing techniques like PacBio. As a result, genes in PacBio raw reads have only around 80% sequence similarity to the assembled gene. In ad-

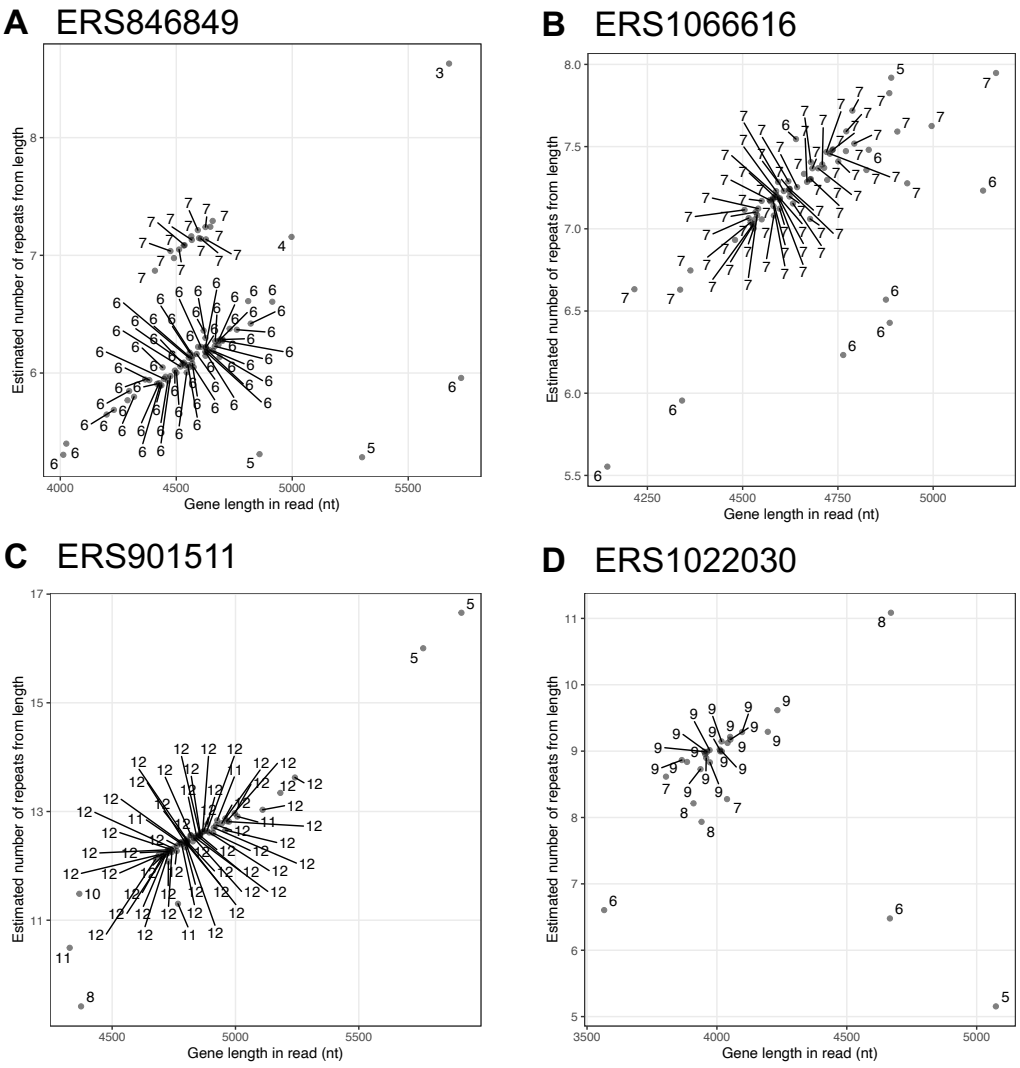

FIGURE 3.8 Analysis of repeat number variation in PacBio raw reads. Scatter plot of estimated number of tandem repeats in PacBio raw reads of two Periscope genes in 4 different bacterial genomes: SasG in *S. aureus* NCTC10443 (A) and NCTC13758 (B) strains, with 7 and 8 repeats respectively; and Sgo_0707 in *S. gordonii* NCTC7865 (C) and NCTC3165 (D) strains, with 12 and 9 repeats respectively. Number of repeats are estimated using three measures: number of nhmmer domain hits (labelled as numbers), length of the region covered by nhmmer domain hits (y-axis), and gene length in the read (x-axis).

dition, other larger insertions of several hundreds of bases were observed in the repeating region of the genes in some of the reads, as in the examples shown in Figure 3.9, which complicated further the estimation and comparison of repeat numbers. Such large insertions are not common in PacBio reads, so their high frequency (around 1 in 10) in the repeating region of Periscope genes is intriguing and may be caused by their high repeat similarity and genomic instability.

I have therefore not been able to find evidence for repeat number variability within a bacterial strain in any Periscope protein. The gene read coverage in the NCTC3000 dataset, that is number of raw reads fully enclosing the gene of interest, was only on the order of 10-100. Recombination events that cause the repeat number changes are therefore not likely to occur at frequencies higher than 1 in 100.

3.3.2 Between strain variability

As no variation within strains could be detected, I analysed the repeat number variation between homologous Periscope proteins from different bacterial strains. I selected well-known Periscope proteins found in a large number of NCTC strain genomes: SasG, SasY, Rib, Sgo_0707, CshA and CdrA. I extracted the assembled protein sequences from each genome and counted the number of repeats in each protein as the number of hits to the corresponding repeat sequence using HMMER. Since assembled genes are of high quality and highly similar to each other, the repeat numbers were reliable.

I found that the variation in the number of stalk repeats can be extreme in some proteins accounting for more than double the length of the protein (Figure 3.10). The most extreme example is SasY, ranging from 3 to 21 Big_6 repeats in the protein. The surface protein Rib also shows high repeat number variability, between 2 and 16 repeats. Other Periscope proteins show a moderate repeat number range, like CshA, Sgo_0707 and CdrA, although large changes are still observed between closely related strains in the tree. For all proteins, the number of repeats is only weakly correlated with genome similarity, suggesting a rapid evolutionary rate of repeat number change in Periscope genes.

A first observation on the repeat number variations suggested that the sequence identity of repeats may play a role in the magnitude of repeat number changes, as repeats in Rib and SasY are 100% identical, whereas CshA and CdrA

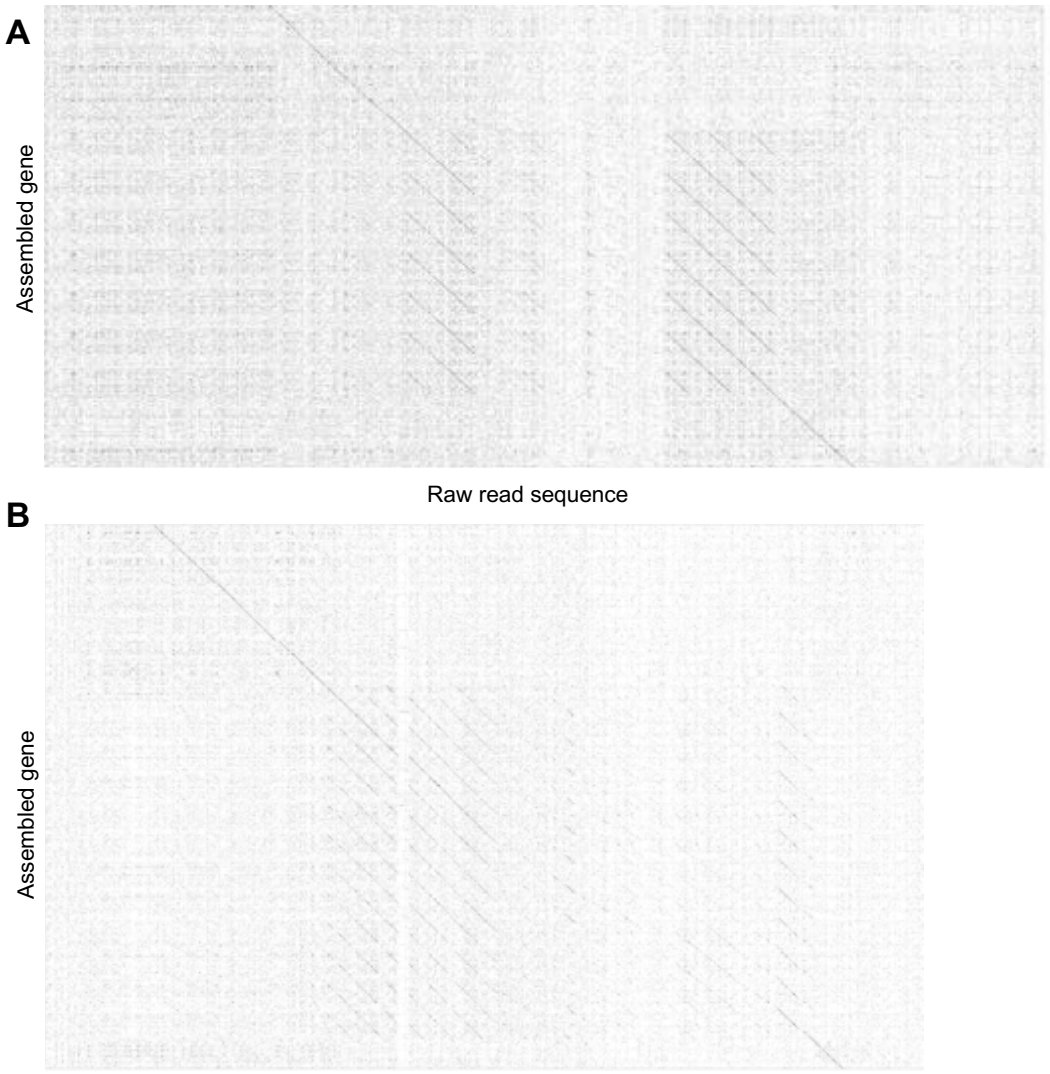

FIGURE 3.9 Dotplot comparison of assembled genes and PacBio raw reads. A) SasG gene from *S. aureus* NCTC10443 and PacBio raw read number 9608, and B) Sgo_0707 gene from *S. gordonii* NCTC7865 and read number 13394. Raw reads cover the entire gene length and have around 80% sequence identity to the assembled gene. Both reads show large insertions in the repeating region of the gene (parallel off-diagonal lines) that do not match any other gene region and are therefore likely sequencing artifacts.

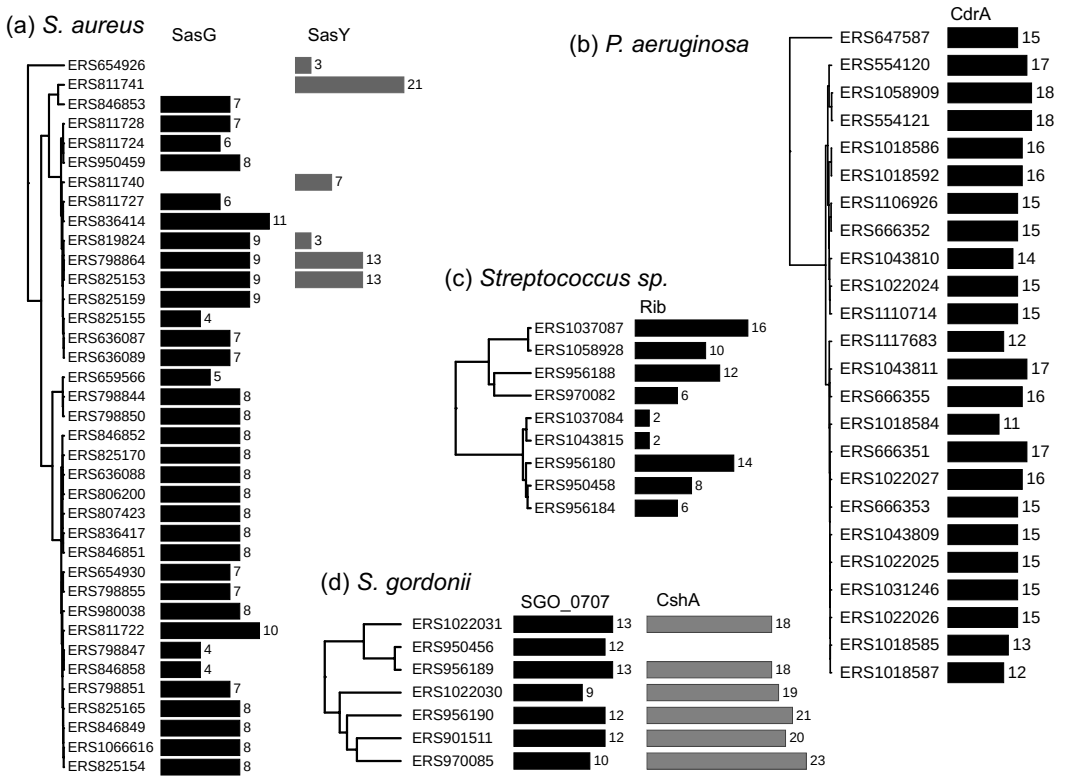

FIGURE 3.10 Variation of stalk domain repeat numbers in Periscope proteins. Phylogenetic trees of *S. aureus* (a), *P. aeruginosa* (b), various Streptococcal species including *S. agalactiae* and *S. pyogenes* (c), and *S. gordonii* (d) genomes in the NCTC3000 collection, mapped to the number of stalk domain repeats in Periscope genes: respectively, SasG gene with double domain G5 and E repeats, SasY gene with Big_6 domain repeats, surface protein CdrA with MBG_2 domain repeats, surface protein Rib with Rib domain repeats, SGO_0707 homolog containing SHIRT domain repeats, and surface adhesin CshA with 100-residue CshA_repeat domains. Phylogenetic trees of bacterial strains are based on genomic comparisons as described in the Methods section and visualised using iTOL (Letunic and Bork, 2019). Figure adapted from Whelan *et al.* (2020).

showed some sequence diversity. Overall, I found the magnitude of the repeat number variation is positively correlated with the percentage of DNA identity of repeats, and that the most extreme repeat numbers and length variation are found in proteins with the highest repeat similarity (Figure 3.11).

3.3.3 Evolution of stalk domain repeat regions

To further study the repeat number variations observed in Periscope proteins, I compared the domain similarity across homologous proteins of different strains within each of the Periscope groups. The analysis for Sgo_0707 homologs shows that the repeat number differences among proteins are mainly found in two regions of the highest repeat DNA identity (Figure 3.12), in line with our previous observations on the role of DNA identity in repeat number variation. In other proteins, such as Rib, the domains in the repeating region are almost 100% identical to each other, making it impossible to determine domain duplication events and locations using the sequence alone.

I also note that domains at the termini of repetitive regions have a lower sequence identity to the rest of the tandem domains, but are highly similar across the terminal domains in other homologous proteins. This observation occurs in almost all Periscope proteins found in this study, including Rib, CdrA, SasG, and CshA. This further suggests that domain duplications, repeat expansions and contractions occur at the middle of the tandem repeat region.

3.4 Discussion

In this chapter, I have defined a new class of bacterial surface proteins named Periscope proteins, which are implicated in mediating interactions between bacterial cells and their environment. I described several features of these proteins, such as the presence of highly similar tandem globular domain repeats, named stalk domains, which form rigid rod-like structures that project the protein out of the bacterial surface. Periscope proteins further exhibit length variation via stalk domain repeat number changes, modulating their surface exposure. I used these unique feature of Periscope proteins to discover new examples in a specialised dataset of bacterial genomes and study the repeat number variation within and between strains.

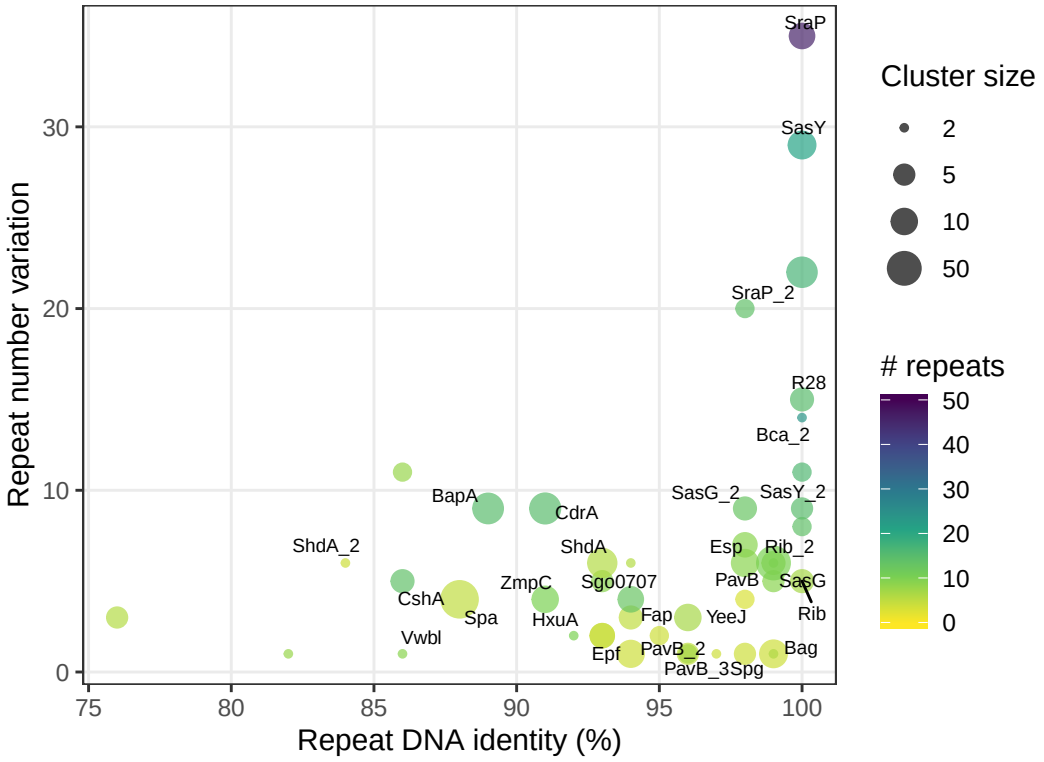

FIGURE 3.11 Repeat number variation in Periscope proteins as a function of repeat sequence identity. The sequence identity of tandem repeats in Periscope genes is plotted against the variation in repeat number observed for each Periscope protein cluster. The repeat number variation is calculated as the difference between the maximum and minimum observed repeat numbers in proteins within each cluster. The maximum number of repeats is shown as a viridis color scale. The repeat DNA identity is calculated as the maximum repeat sequence identity calculated by T-REKS across proteins in each cluster. The number of proteins in each cluster is shown as point size. Names for the most relevant Periscope protein clusters are shown as labels. Figure adapted from Whelan *et al.* (2020).

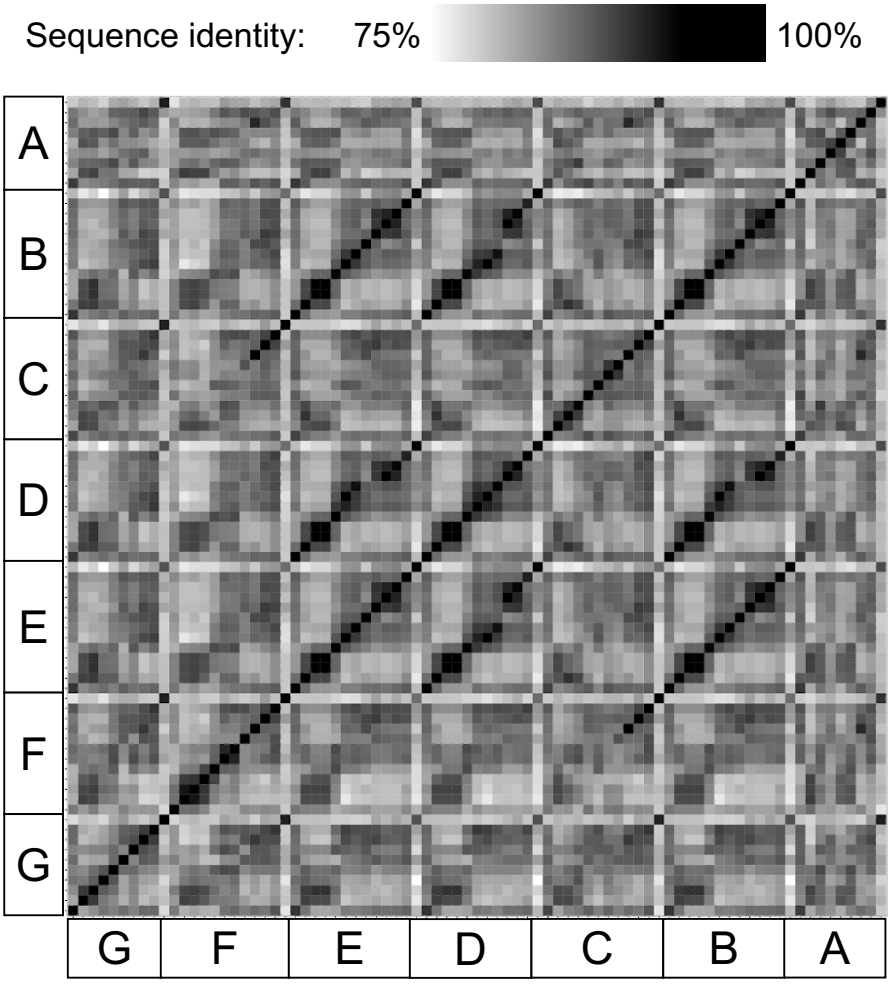

FIGURE 3.12 Similarity matrix of SHIRT domains in genes coding for Sgo_0707 proteins. All-by-all domain DNA similarity matrix of SHIRT domains (sequential order) in seven Sgo_0707 proteins from different NCTC3000 strains (A-G). Main diagonal boxes correspond to domains within the same gene, while off-diagonal entries correspond to comparisons of domains in different genes. Domain similarities, insertions and deletions in the matrix are interpreted in a similar way to sequence dotplots, but entries represent DNA similarity of domains instead of amino acid similarity. For example, genes of proteins B and D appear to be close homologs, as their center domain diagonal is highly similar, but domain deletion in protein D can be observed around the 9th domain. The first domain in the tandem domain repeat region of each protein is more dissimilar to the rest, so a grid pattern that separates domain regions is formed. The pairwise sequence identity between domains is calculated from the Pfam sequence alignment (omitting gaps).

I have computationally identified over 50 groups of bacterial proteins that can be potentially classified as Periscope proteins, which are widespread across bacterial species and have diverse domain compositions. Over half of these proteins are uncharacterised, while others are well known bacterial surface proteins but had not been functionally and structurally associated together before. Repeat sequences in these potential Periscope proteins further confirm that stalk domains fold into globular domain units, providing a mechanism to modulate the protein length by changing the number of units (or domains) in the stalk.

I further observed a strong amino acid composition bias in the stalk domain repeat regions of Periscope proteins. This bias is specific to the tandem domain repeats, and it is not shared among proteins: despite few amino acids are highly enriched in some proteins (such as Thr and Pro), other proteins are depleted in these same amino acids and show enrichment for others (such as Gly or Leu in CdrA). It appears that the only common denominator is the presence of a composition bias, but no general principles about the type of bias or its functional role could be found.

I could not find evidence of repeat number variability within individual genomes using raw sequencing PacBio reads, but I found a high repeat variability in Periscope proteins from different bacterial strains. The uneven distribution of the number of repeats across the strain phylogenetic trees suggests the presence of a molecular mechanism that enables rapid changes in repeat numbers at the genomic level.

A further source of variability in Periscope proteins is the interchange of stalk domains between two different proteins, while maintaining a similar N-terminal adhesive domain. This was initially observed in proteins SasY and Rib, which share an homologous N-terminal adhesin AlphaC-like domain pair, but their stalks are formed by Big_6 domains in SasY and Rib domains in the surface protein Rib. Other examples of proteins with the same stalk domain, but different N-terminal domains, are found in Table 3.1, but these stalk domain exchanges have not been studied systematically in this chapter.

The major limitation for the study of Periscope proteins has been the availability of high quality genomic data. The small number of bacterial strains from a limited subset of species in the NCTC3000 genomes hindered the discovery of potential Periscope proteins and the study of their repeat number variability.

Furthermore, the sequencing coverage of PacBio long reads on the individual Periscope genes was not high enough to study repeat number variations at the single strain (or genome) resolution. Samples with higher sequencing coverages will be needed to observe these types of variations, and it is likely that targeted experiments are required to increase read coverage levels of Periscope genes to a sufficient level in order to observe changes in the repeat number within the same genome.

Two open research questions that I have not addressed in this study are the relation between the Periscope protein repeat numbers and the bacterial cell envelope thickness, which is also known to vary across bacterial strains, and the density of these proteins at the bacterial surface. Both the bacterial envelope thickness and protein expression levels are hard to predict *in silico* from genome sequences alone, but current state of the art imaging techniques have shown promising results to investigate them.

The repeating nature of these surface exposed proteins further opens questions about their immunogenicity: repeats would be difficult to mutate in case the immune system is able to recognise them, for example through specific antibodies or sequence motifs. Several studies have reported these types of immune selective pressure in Periscope proteins like CshA (Elliott *et al.*, 2003). Repeat number variation and sequence composition biases could be explained as a consequence of immune escape mechanisms.

This study sets the foundation to understand the nature and properties of Periscope proteins, a subset of bacterial surface proteins with special interest due to their involvement in important bacterial functions such as host invasion and biofilm formation. These findings have also opened several research avenues for further experimental and computational investigations.

3.5 Methods

3.5.1 Extraction of NCTC3000 proteins

For each bacterial strain in the NCTC3000 dataset, the assembled genome sequence in FASTA format and its gene annotations in GFF format were downloaded from the project's FTP site[3] as of October 2019. A total of 734 bacterial genomes corresponding to 207 different bacterial species had gene annotations at the time of this study. Using gene boundaries from the GFF files, a total of 2,579,577 proteins and their associated coding transcripts were extracted. New unique identifiers containing information on the genome, contig, strand orientation and location were assigned to proteins and coding transcripts, to be used internally for sequence searches and other analyses.

3.5.2 Clustering of repeats and full protein sequences

Sequences of repeats were clustered by sequence similarity using all against all BLASTp (Altschul *et al.*, 1997) searches. Significant BLASTp hits were used to build a sequence similarity network with the `igraph` package in `R` (Figure 3.4). Finally, a bit score threshold of 30 units was applied to the network to extract clusters of connected components from the graph. Additionally, Pfam families were assigned to a representative sequence from each cluster, chosen at random, searching the Pfam version 32.0 HMM library with HMMER.

Full length proteins were clustered with a similar approach using all against all BLASTp search and extracting connected components, but using an additional filter of 90% sequence identity threshold.

3.5.3 Analysis of PacBio sequencing raw reads

Selection of raw reads

PacBio raw sequencing reads were downloaded from ENA in H5 format and converted to plain FASTA sequences using the DEXTRACTOR tool [4]. The assembled sequence of the gene of interest was searched against the raw reads using

[3] `ftp://ftp.sanger.ac.uk/pub/project/pathogens/NCTC3000`
[4] `https://github.com/thegenemyers/DEXTRACTOR`

BLASTn (Alford *et al.*, 2017). The reads that covered entirely the gene were selected, meaning that the sequence range of the BLAST hits covered both N- and C-terminal regions.

Repeat number counts

A sequence alignment of the repeats in the Periscope gene of interest was generated by running T-REKS (Jorda and Kajava, 2009) on the assembled gene sequence. The nhmmer program of the HMMER tool (Wheeler and Eddy, 2013) was then used to search for matches of the repeat in raw reads covering the full gene. Checks to ensure nhmmer hits were complete (coverage higher than 95%) and of high similarity to the repeat sequence (e-value lower than 10^{-10}) were introduced in order to filter out false positives. Additionally, the number of repeats was estimated solely from the length of the repeating region using the lowest (start) and highest (end) nhmmer hit indices as following:

$$count = \frac{end - start}{repeat\ length} \tag{3.1}$$

The two repeat counts based on the number of nhmmer hits and repeat region length were used in combination to search for genes with a different number of repeats to the assembled gene.

3.5.4 Phylogenetic trees of NCTC3000 strains

Phylogenetic trees for strains in Figure 3.10 were created by comparing all genes in a genome to all genes in another genome using BLASTn. Hits with a difference in sequence length over 100 residues and coverage below 50% were removed in order to filter out partial matches. For each gene, the highest scoring pair was chosen in case of multiple hits. For each pair of genomes, a similarity score was calculated as the weighted average of sequence identities of all pairs of homologous genes, as reported by BLASTn. A dendrogram of strains was constructed by hierarchical clustering of the pairwise sequence identity matrix using the `hclust` function from `stats` package in `R`. Phylogenetic trees were visualised in the online tool iTOL (Letunic and Bork, 2019).

3.5.5 Calculation of sequence bias and skew

The relative entropy (also called Kullback–Leibler divergence, D_{KL}) is a measure of the difference between two probability distributions, and can be used to measure the composition bias of a sequence similarly to the Shannon entropy. It is more intuitive — the lowest value of 0 means that the two distributions are identical and higher values represent increasing bias in the composition — and more flexible, since the background probability distribution is a parameter that can be chosen.

The relative entropy H_R of a protein sequence S is computed from the individual observed f_i and background b_i frequencies of the 20 amino acids as following:

$$H_R(S) = \sum_{i}^{20} f_i \log_2 \left(\frac{f_i}{b_i} \right) \tag{3.2}$$

Here I used a uniform background amino acid distribution with $b_i = 0.05$ for all amino acids.

Nucleotide composition metrics are calculated as following, where A, T, C, G are the raw counts of each nucleotide in the sequence.

$$Asymmetry(A,T) = \frac{A-T}{A+T+G+C} \tag{3.3}$$

$$Skew(A,T) = \frac{A-T}{A+T} \tag{3.4}$$

The set of random genes was constructed by randomly selecting ten genes from each NCTC genome that contained at least one Periscope protein.

Chapter 4

Tandem domain repeat structures

> *"These results imply that the degree of success to be expected in predicting the structure of a protein from its sequence using the known structure of an homologous protein, depends upon the extent of the sequence identity."*
>
> - Chothia and Lesk (1986): "The relation between the divergence of sequence and structure in proteins."

In this chapter, I explore the properties and evolution of individual tandem domain repeat structures from bacterial surface proteins experimentally determined by our collaborators at the University of York, and others available in the Protein Data Bank (PDB). I further suggest domain constructs for experimental characterisation and structure determination to our collaborators, including a subset of ten domains for automated structure determination at the ESRF Synchrotron in collaboration with Matthew Bowler.

The results on the Rib and SHIRT domains have appeared in the following two publications, respectively: Whelan *et al.* (2019) and Whelan *et al.* (2020). I am a co-first author on both articles and actively participated in data analysis, figure generation and manuscript writing. I have not participated in wet-lab experiments or structure determination, which was done independently by members of Jennifer Potts' research group at the University of York and Matthew Bowler at EMBL Grenoble.

4.1 Introduction

Protein structures are key to understand biological functions: they illuminate molecular mechanisms with atomic detail, they reveal specific interactions between components in signalling networks and, since protein structure is more conserved than sequence, they uncover remote evolutionary relationships between proteins (Chothia and Lesk, 1986).

The research lab of Prof. Jennifer Potts at the University of York has recently made important discoveries on tandem domain repeats in bacterial surface proteins. They have been studying experimentally the biophysical properties of domain repeats in isolation and determined experimentally several structures, shown in Figure 4.1, including the double domain repeat in SasG (Gruszka *et al.*, 2012), the Rib domain (Whelan *et al.*, 2019), the SHIRT domain (Whelan *et al.*, 2020), and a domain of unknown function (Pfam:DUF1542) with a three-helical bundle fold related to B and GA bacterial surface domains (structure not yet released). Simultaneously, other research labs have been working on further repetitive bacterial surface proteins, such as the CdrA protein from *Pseudomonas aeruginosa* (Melia *et al.*, 2021), which contains domain repeats from a new domain superfamily named MBG and an available homologous domain structure (Figure 4.1D), and the CshA protein from *Streptococcus gordonii* (Back *et al.*, 2020).

These domain structures have been key to discover that long tandem repeats commonly found in bacterial surface proteins fold into independent globular domains, and to formulate hypotheses about their function as surface-projecting stalks. These structures further show features unexpected in small stable globular domains, possibly a consequence of unusual selective pressures, and provide an interesting set to study rare evolutionary events such as domain atrophy (Prakash and Bateman, 2015).

In close collaboration with structural biology labs, I have analysed several experimental structures of tandem domain repeats in bacterial surface proteins in order to gain insights into their properties and evolution, suggesting further domain constructs for experimental characterisation and structure determination. In this chapter, I describe the main results of this work.

In the next introductory subsections, I briefly review protein structure determination and prediction techniques, and introduce the concept of domain atrophy, a rare evolutionary event that has been frequently observed in structures of

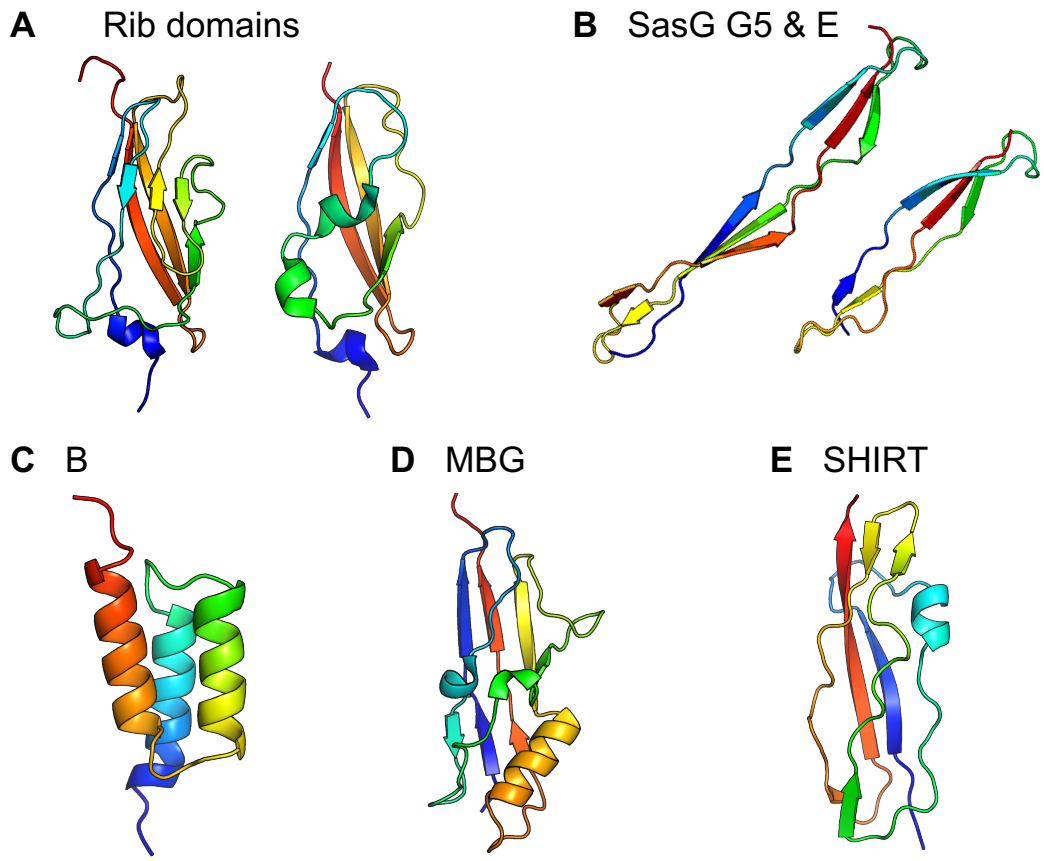

FIGURE 4.1 Experimental structures of tandem domain repeats in bacterial surface proteins: A) Rib domains solved by Whelan *et al.* (2019): left, Rib long (PDB:6S5W) and right, Rib standard (PDB:6SX1); B) tandem G5 (left) and E (right) domains from the protein SasG (PDB:3TIP) solved by Gruszka *et al.* (2012); C) immunoglobulin-binding domain (B domain) from *Staphylococcus aureus* virulence factor protein A (PDB:1H0T); D) first structure of an MBG domain (PDB:4NG0), solved by Etzold *et al.* (2014); and E) SHIRT domain from *Streptococcus gordonii* (PDB:7AVK) solved by Whelan *et al.* (2020). Structures are shown as cartoon representations and rainbow colored from N-terminal (blue) to C-terminal (red), and shown in a similar orientation with the N-terminus pointing down. Structures visualised using PyMol.

tandem domain repeats.

4.1.1 Protein structure determination

There are several experimental techniques to determine the structures of proteins at varying levels of resolution; the most popular techniques are X-Ray crystallography, Cryo-electron microscopy (Cryo-EM) and Nuclear Magnetic Resonance (NMR).

In X-Ray crystallography, proteins are first crystallised from highly pure and concentrated samples in a trial and error process. Protein crystals are then illuminated with a bright beam of incident X-rays, usually from a synchrotron, and the resulting X-Ray diffraction pattern is used to reconstruct a three-dimensional electron density map of the protein (Smyth and Martin, 2000). The chain of amino acids is then threaded into the electron density map to generate an atomic model of the protein structure. X-Ray crystallography has been used to determine structures of small domains and big protein complexes alike. The protein crystallisation step is very unpredictable and commonly the bottleneck of structure determination projects, although modern high throughput crystallisation screens have dramatically improved the success in recent years and are very successful for small stable globular domains such as the ones observed in tandem domain repeats. X-Ray crystallography is the most common structure determination technique, accounting for almost 90% of structures in the Protein Data Bank (PDB) — the reference database of experimentally determined protein structures (Berman *et al.*, 2000) — at the time of this study, although it has been losing its historical hegemony to Cryo-EM in recent years.

In Cryo-EM, thousands of single particle images of protein molecules are picked from electron micrographs of purified protein samples in cryogenic conditions, and combined using sophisticated image processing software to reconstruct the three-dimensional electron density map of the protein (Jonic and Vénien-Bryan, 2009). If the electron density map reaches atomic resolutions (lower than 5Å), atomic coordinates can be fitted in a similar way to X-ray crystallography to generate an atomic model of the protein structure. Cryo-EM has recently experienced a resolution revolution and a surge in popularity and applications of the technique. Sample preparation is simpler than for X-Ray crystallography and proteins largely remain in native conditions, but technological limits make the

technique only applicable to large proteins or protein complexes above 100 kDa or 1,000 amino acids and with limited flexibility.

NMR is a spectroscopic technique that enables the inference of interatomic distance bounds in proteins, which can be further used as restraints in specialised modelling frameworks to generate atomic models of proteins (Wüthrich, 2001). NMR requires sophisticated sample preparation techniques and is only applicable to small domains, so it is sparsely used as a structure determination technique. It has other advantages for the study of proteins such as measuring dynamics and conformational ensembles, or studying intrinsically disordered proteins.

Other experimental techniques permit the study of protein structures at lower resolution. Cryo-electron tomography is used to study biological macromolecules *in situ* within cells (Schaffer *et al.*, 2019), and offers many advantages for cell-biology and functional studies. Melia *et al.* (2021) used the technique recently to observe bacterial fibrillar adhesins at the surface of *Pseudomonas aeruginosa* and their interactions at the molecular level. Small angle X-Ray Scattering (SAXS), on the other hand, enables the study of protein molecules and their complexes in solution as low resolution shapes, providing information on their overall dimensions as elongated or spherical forms (Gräwert and Svergun, 2020). In repetitive surface proteins, SAXS data has been used as evidence of their fibrillar elongated shapes (Whelan *et al.*, 2019). These techniques are not yet capable of reaching atomic resolutions below 5Å and have therefore not yet been used for protein structure determination.

Solving structures of tandem domain repeats experimentally presents additional challenges, such as defining domain boundaries accurately, characterising the stability of domains in isolation and the additional flexibility caused by inter-domain linkers in constructs of tandem domains. Incomplete or wrong domain boundaries cause low stabilities and reduce the crystallisation success, as demonstrated by early attempts to determine the structure of the SHIRT domain (Whelan *et al.*, 2020). Studying the sequences of domain repeats in their evolutionary context, comparing them to other domains in the same family and other related families with known structure, is crucial to correctly define domain boundaries in tandem repeats.

4.1.2 Protein structure prediction

Determining protein structures experimentally is a costly and time consuming task, and the number of natural proteins far outnumbers our current experimental capabilities; computational protein structure prediction techniques are therefore fundamental for many research projects in biology.

To our advantage, protein structures are more conserved than their amino acid sequences, so known experimental structures can be used to infer the structures of close homologs with good accuracy, an approach known as homology-modelling. However, experimental structures are still not enough to homology-model all known proteins because they only cover about two thirds of the protein sequence space, and the accuracy of these methods decreases with protein size and sequence divergence.

Another group of techniques aim to model protein structures *de novo* directly from their amino acid sequence. These methods use models trained on protein experimental structures in the PDB and evolutionary features inferred from multiple sequence alignments of related proteins, such as residue conservation and correlated mutations (Marks *et al.*, 2011), and have gained a lot of popularity in recent years due to their unprecedented accuracy at the Critical Assessment of protein Structure Prediction (CASP) experiment, a biannual international challenge for blind protein structure prediction (Moult *et al.*, 1995). Their success is in large part thanks to the use of powerful deep learning models that are capable of learning more meaningful representations of protein sequences and structures. The AlphaFold system (Senior *et al.*, 2020) raised to the top of the rankings in CASP14 (2020) achieving atomic accuracy comparable to models derived from experimental data.

The current computational cost and accuracy of structure prediction techniques enables to reliably model protein structures at large-scale, for example achieving near-complete proteome levels (Greener *et al.*, 2020). The transform-restrained Rosetta (trRosetta) (J. Yang *et al.*, 2020) is a *de novo* structure prediction method from David Baker's lab at the University of Washington based on deep learning that consistently predicted accurate models in CASP14 and ranked second only behind AlphaFold. The Baker lab has recently generated trRosetta models for nearly all families in Pfam, dramatically increasing the structural coverage of protein families (unpublished internal collaboration with Pfam, data not

yet available). This data offers a unique opportunity to complement the analyses of experimental structures of tandem domain repeats with computational models.

4.1.3 Domain atrophy

Prakash and Bateman (2015) observed a small number of large-scale deletions of core structural elements in protein domains, and hypothesised that these create shorter versions of the original domain structure. They named these evolutionary events "domain atrophy" and found that they are extremely rare across all known protein sequences, including stable ancient folds such as TIM-barrels and Rossmann folds. Although structural losses do not affect functional sites, they observed compensatory mutations and other consequences such as protein dimerisation.

At the time, Prakash and Bateman (2015) did not find many examples of domain atrophy with structural evidence, meaning that structures for the atrophied and complete domain pairs are available. Since then, experimental structures of tandem domain repeats have shown remarkable structural malleability and surprisingly evident examples of domain atrophy. For example, E domains from the surface protein SasG have evolved from their adjacent G5 domain through two large structural deletions that shortened the main β-sheet structure of the domain by about one third. This, and other domain atrophy examples, are analysed in the following section.

4.2 Analysis of domain structure evolution

4.2.1 The G5 and E domains in SasG

The *Staphylococcus aureus* surface protein G (SasG) is a large multidomain protein attached at the cell-wall that forms rod-like fibrils and is implicated in cell adhesion and biofilm formation (Corrigan *et al.*, 2007). The central region of the protein is formed by large repeats of 128 amino acids that fold into a domain pair of G5 (Pfam:PF07501) and E domains (Pfam:PF17041).

The G5 and E domain repeats were one of the first structures of bacterial surface tandem domain repeats experimentally solved by the group of Jennifer Potts (Gruszka *et al.*, 2012). G5 domains fold into long triple-stranded single-layer β-sheets (Figure 4.1B) and are widespread in bacterial extracellular proteins

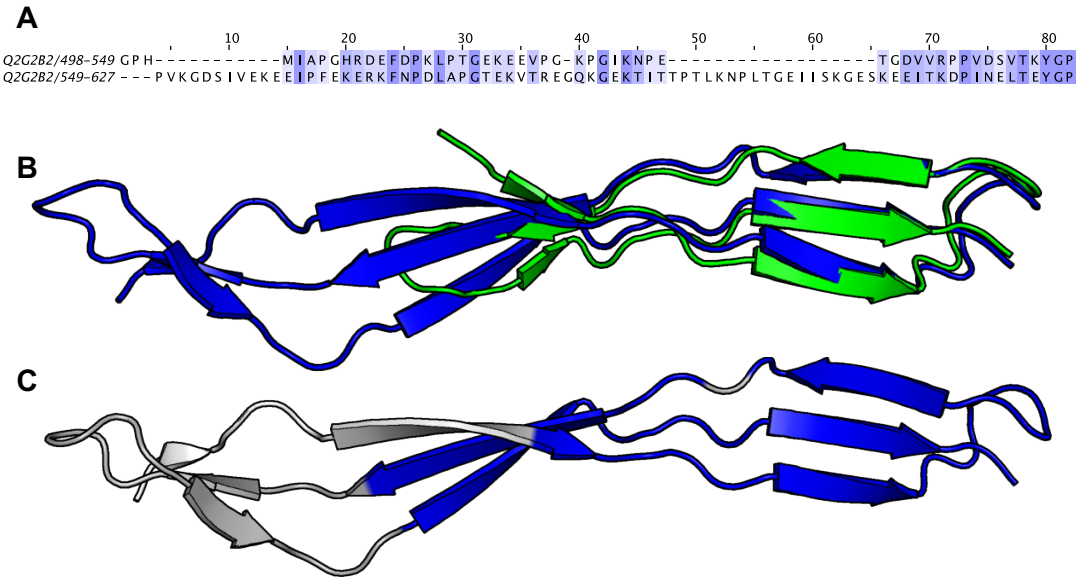

FIGURE 4.2 Alignment of the G5 and E domains in SasG. A) Structure-based sequence alignment of the E (top) and G5 (bottom) domains colored by sequence similarity (BLOSUM62), with a total 23% sequence identity (omitting gaps). B) Structural superposition of the G5 domain (blue) and the E domain (green) based on the sequence alignment with 1.7Å C_α RMSD. C) Structure of the G5 domain in blue, with structural deletions corresponding to gaps in the alignment to the E domain highlighted in grey. Sequence alignments visualized using Jalview (Waterhouse *et al.*, 2009), and structures using PyMol.

with over 20 thousand domains in UniProt across different species. E domains, on the other hand, are only found adjacent to other G5 domains in SasG and another related protein of *Staphylococcus epidermidis*. Gruszka *et al.* (2016) found that the E domain is not stable in isolation, but that G5 and E domain pairs fold cooperatively into very stable structures, despite being predicted to be disordered.

Close inspection of the E domain structure reveals that it folds into a short triple-stranded single-layer β-sheet similar to G5 domains (Figure 4.2B). Based on their sequence and structural similarity (23% sequence identity and 1.7Å C_α RMSD), Alex and I hypothesised that the two domains are homologous, and that the E domain has evolved from its adjacent G5 domain through two large structural deletions at the N-termini and around the β-hairpin formed by the second and third strands (Figure 4.2C).

4.2.2 The Rib domains

Rib domains are mostly found as nearly identical tandem repeats in a group of streptococcal cell surface proteins named α-like proteins (Wästfelt *et al.*, 1996), which have the typical Periscope architecture of stalk domain repeats described in the previous chapter (Figure 4.3). They have been extensively studied for many years due to their demonstrated implications in strain pathogenicity, and have been of special interest to microbiologists (Gravekamp *et al.*, 1996).

The first structure of a Rib domain repeat was recently determined using X-Ray crystallography in the lab of Jennifer Potts and initially appeared to be a novel β-sandwich type fold (Whelan *et al.*, 2019). More detailed structure analysis revealed that the Rib domain is actually remotely related to Immunoglobulin domains but has a large deletion in place of the D and E strands characteristic of the B-E-D β-sheet (Figure 4.4C).

Alex and I observed that some domains in the Rib family (Pfam:PF08428) contained long sequence insertions in the position of the missing DE strands (Figure 4.3C). We hypothesised that these domains would retain the full Immunoglobulin fold, representing something akin to the ancestral form of the Rib domain, and suggested sequences of these long Rib domains from a surface protein of *Lactobacillus acidophilus* (UniProt:Q5FIM8), containing both long and short Rib domains with 45% sequence identity, to our collaborators at the University of York, who successfully determined two domain structures experimentally. As predicted, the missing D and E strands in Rib are present in the longer version of the domain (Figure 4.4B), which comprises a full Ig-like fold. A separate family for these longer domains was included in Pfam, named RibLong (Pfam:PF18957).

I further identified by structural similarity another two domains related to Rib in the *S. gordonii* GspB protein (Pyburn *et al.*, 2011) and the *S. sanguinis* SrpA adhesin (Bensing *et al.*, 2016). The topology of these domains is very similar to the Rib domain, with 2.95Å and 2.4Å C_α RMSD respectively (Figure 4.4D). The antiparallel G-F-C β-sheet at the C-terminus of the domains is smaller but highly conserved, but a long α-helix now replaces the B strand in the Rib structure, covering the space of the atrophied D-E strand pair. Although their sequence is more divergent than RibLong domains, these domains share the highly conserved YPDxxxD motif with Rib domains (Figure 4.4A), highlighting the importance of the motif in stabilising the F-G β-hairpin and the homology of these domains to

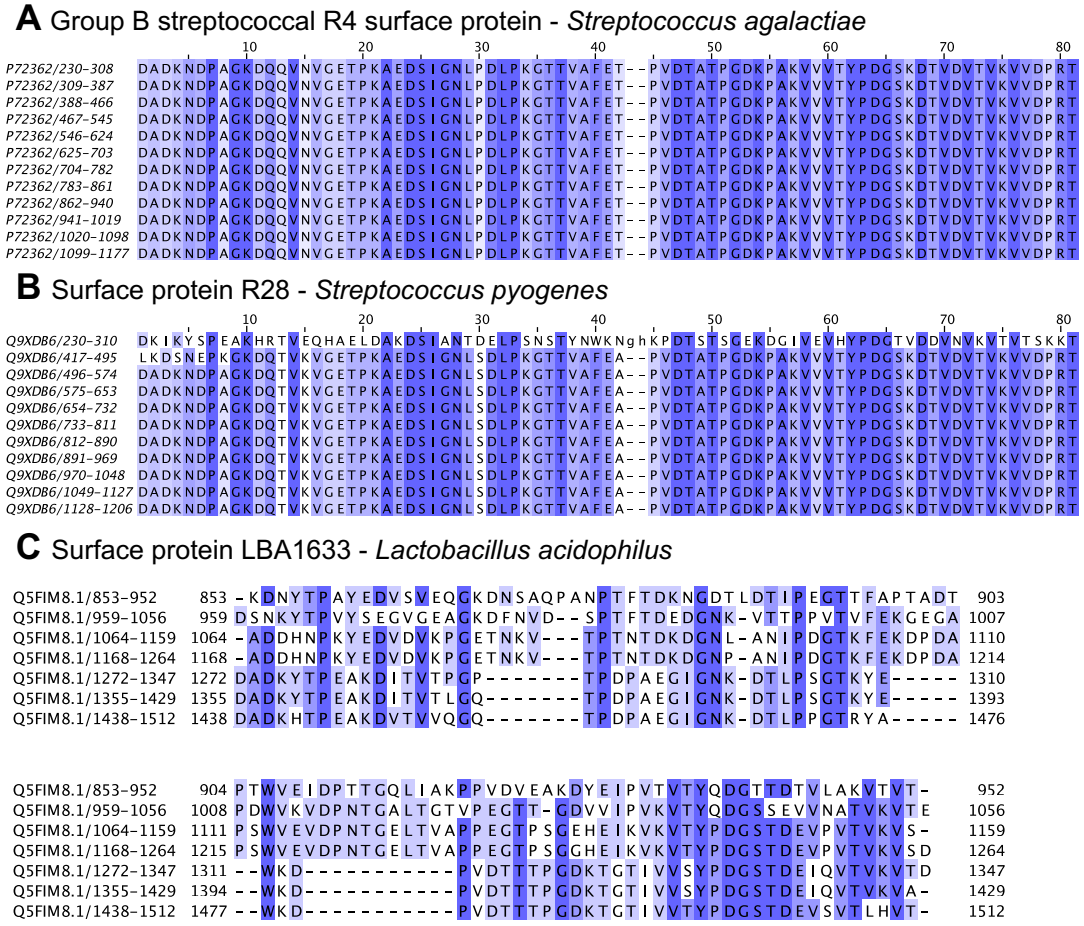

FIGURE 4.3 Alignment of tandem Rib domain repeats in surface proteins from *S. agalactiae* (A), *S. pyogenes* (B) and *L. acidophilus* (C). Domains in R4 (A) and R28 (B) surface proteins are part of the same alignment for comparison purposes. Domains in R4 are identical to each other and share 95% sequence identity to domains in R28, except for the two most N-terminal domains. The three C-terminal tandem Rib domains in LBA1633 (C) are more divergent and longer than the other, with two main insertion sites. Alignments visualized using Jalview (Waterhouse *et al.*, 2009) and colored by sequence identity. Figure adapted from Whelan *et al.* (2019).

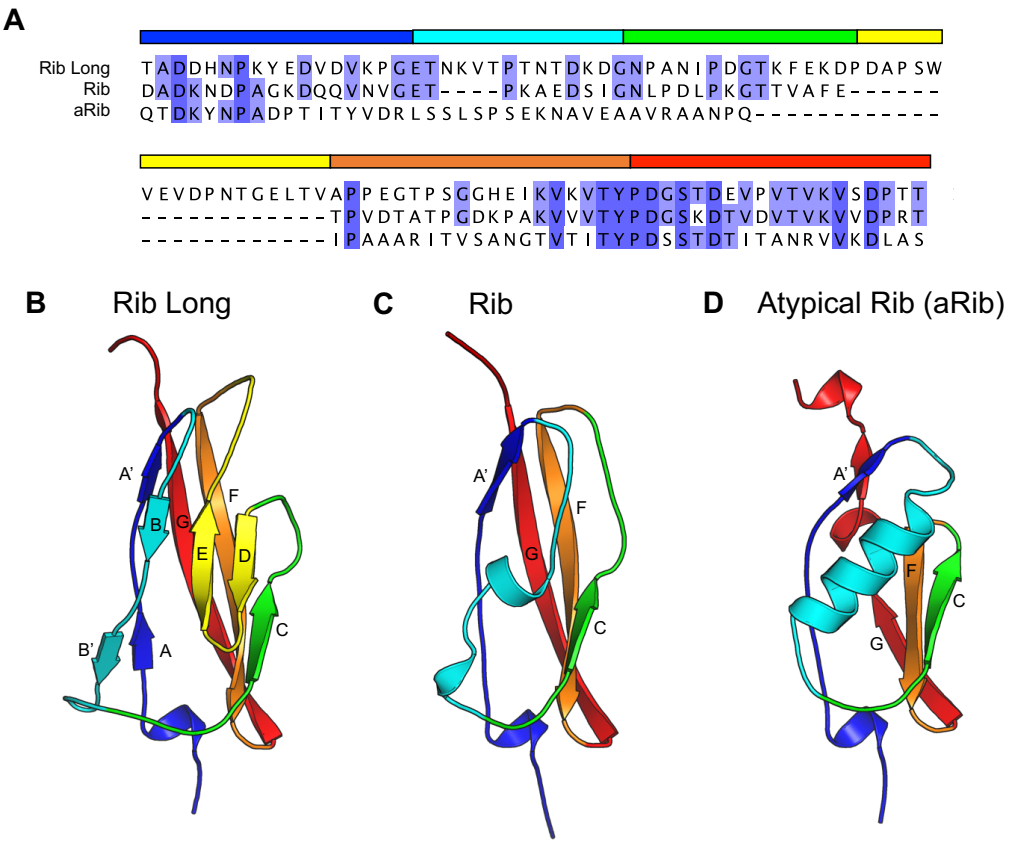

FIGURE 4.4 Comparison of the sequence (A) and structure of Rib Long (B), Rib (C), and atypical Rib (aRib, D) domains. Sequence alignment is coloured by amino acid sequence identity and shows homologous structural regions as colored blocks on top. Structures of Rib Long (PDB:6S5W), Rib (PDB:6S5X) and aRib (PDB:3QC6) shown in the same orientation (N terminus down, C terminus up) and coloured to illustrate homologous structural regions. Loop conformations of the domains have been smoothed to highlight coarse topological similarities. Sequence alignments visualized using Jalview (Waterhouse *et al.*, 2009), and structures using PyMol. Figure adapted from Whelan *et al.* (2019).

Rib domains. Alex and I built another Pfam family for these domains, previously unclassified, that we named atypical Ribs (aRib: PF18938).

The domain length is mostly conserved within each Rib domain Pfam family, and the average length of each family differs: aRib domains are the shortest with an average around 70 residues, Rib domains have an average around 80 residues and RibLong domains have an average over 90 residues (Figure 4.5). The sequence similarity network of Rib domains further shows that aRib domains are more closely related to Rib domains than to RibLong domains (Figure 4.5). Given the narrow species distribution of the Rib domain family, and compared to the widespread distribution of Immunoglobulin folds, we suggest that RibLong domains are ancestral to all the other Rib domains, which are an evolutionary recent example of domain atrophy with high sequence and structural similarity (Figure 4.4). The aRib domain fold has evolved from Rib domains, and their long α-helix observed is a structural elaboration that covers the exposed domain core caused by the deletion of the DE strands, a compensation mechanism similar to others observed as a consequence of domain atrophy (Prakash and Bateman, 2015).

Despite the large abrupt structural loss in Rib domains, about 20% of the domain compared to the ancestral Rib long domain, they are more thermostable with a measured melting temperature (TM) of 88°C, compared to a TM of 78°C for the Rib long domain (Whelan *et al.*, 2019). These melting temperatures are above the average TM of similar small globular domains.

4.2.3 The SHIRT domain

Sgo_0707 is a surface protein from *S. gordonii* (UniProt:A8AW49) composed of a pair of N-terminal domains involved in collagen binding (Nylander *et al.*, 2013), a C-terminal LPXTG cell-wall attachment motif and multiple previously uncharacterised domain-size tandem repeats in the middle. As the repeats did not match any existing Pfam definition, Alex built a new family and named the putative domain SHIRT (Streptococcal High Identity Repeats in Tandem). The family contains only 174 domain sequences from 86 different proteins, the majority of them as tandem domain repeats (111 domains, 64% in tandem).

The T-REKS method predicts 13 repeats of 84-90 residues with high sequence similarity, above 85% sequence identity, and a repeat frame starting at 460-543. The lab of Jennifer Potts solved experimentally the structure of the

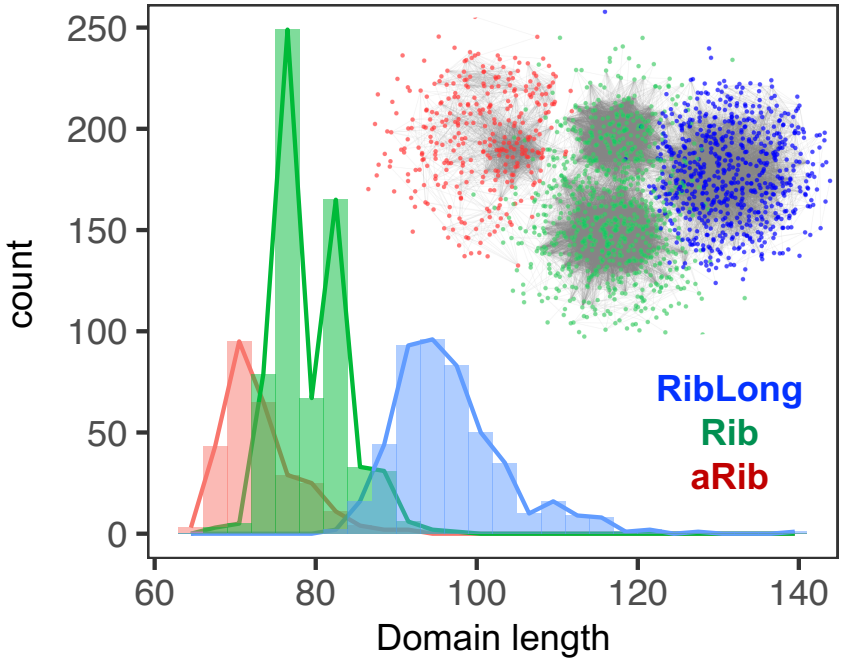

FIGURE 4.5 Comparison of length and sequence similarity of Rib domain families. Distribution of Rib domain lengths from the three Pfam families: RibLong (Pfam:PF18957) in blue, Rib (Pfam:PF08428) in green, and aRib (Pfam:PF18938) in red. Sequence similarity network (SSN) of domains shown as an inset to the plot, with domain sequences as nodes colored by Pfam family and edges representing sequence similarity hits by BLASTp.

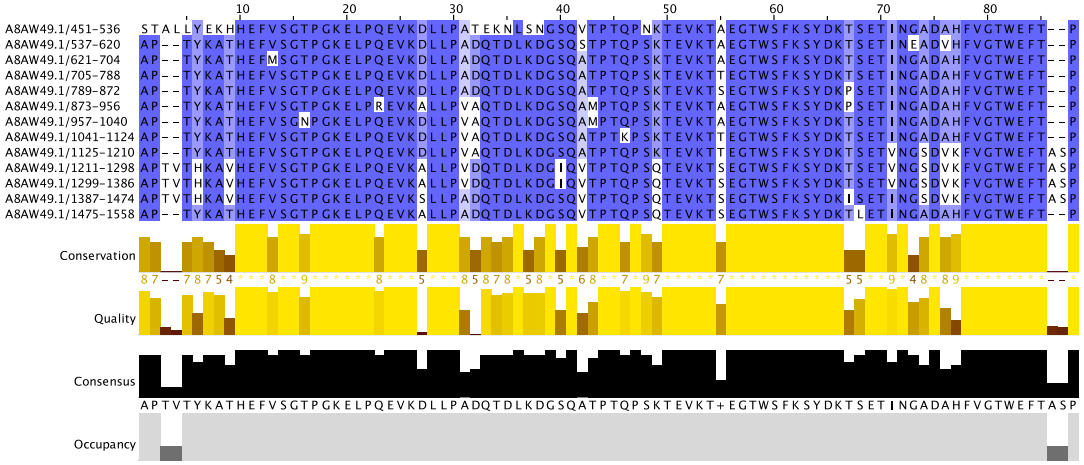

FIGURE 4.6 Alignment of tandem SHIRT domain repeats in the surface protein Sgo_0707 from *Streptococcus gordonii* (UniProt:A8AW49). Alignment of repeats shown in the correct frame according to experimental domain boundaries (537-620), and colored by sequence identity. Annotations for alignment columns at the bottom correspond to conserved physico-chemical properties (Conservation), likelihood of observing mutations (Quality), percentage of the modal residues (Consensus) and number of gaps (Occupancy). Alignment figure and annotations were generated using Jalview (Waterhouse *et al.*, 2009).

SHIRT repeat using X-Ray crystallography, but discovered a significant truncation of the N-terminal β-strand, so they decided to correct the repeat frame by shifting the boundaries seven residues towards the N-terminus, hoping this would complete the fold (Figure 4.6). The new SHIRT experimental structure confirmed the repeat boundaries and showed a remarkably high thermal and chemical stability[1].

The SHIRT domain folds into a β-fold with an unusually low fraction of secondary structures; apart from two long interacting β-strands at the N- and C-termini, the domain only contains short β-strands and a short helical region (Figure 4.7A). The domain is difficult to classify into known domain folds because there are no close structural homologs in the PDB. The B-repeat (Pfam:Flg_new) of *Listeria monocytogenes* internalin B protein (Figure 4.7B), also commonly found in tandem repeats, is the closest domain structure with a C_α RMSD of 3.4Å, and

[1]Data from collaborators in Jennifer Potts lab (University of York) not yet published and not shown here.

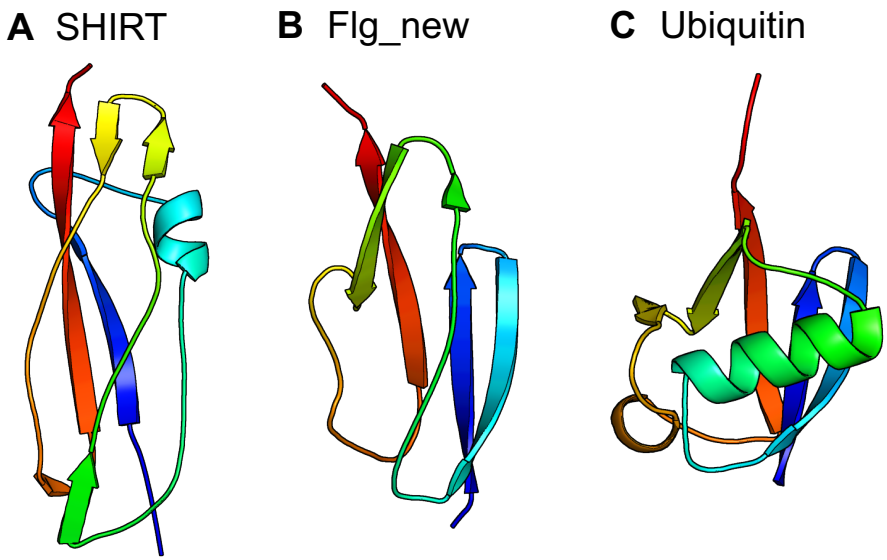

FIGURE 4.7 Comparison of the SHIRT domain structure to other homologous domains with Ubiquitin folds. A) SHIRT domain structure from the Sgo_0707 protein (PDB:7AVK), B) Flg_new domain structure from the internalin B protein (PDB:2Y5P) and C) human Ubiquitin domain (PDB:1UBQ). The domains are distantly related, without detectable sequence similarity and a C_α RMSD between SHIRT and Flg_new domains of 3.4Å, between Flg_new and Ubiquitin of 5.3Å, and between SHIRT and ubiquitin of 6.1Å. Domain structures are rainbow colored from N-terminal (blue) to C-terminal (red) and shown in a similar orientation with N-terminus down and C-termini up. Structures visualised using PyMol with smoothed loop conformations to highlight coarse topological similarities.

it has been classified as a distant relative of ubiquitin domains with β-grasp folds (Ebbes *et al.*, 2011). The Flg_new and SHIRT domains share the same topology, with interacting N- and C-terminal β-strands, but the SHIRT domain is more elongated and its secondary structure is less defined. Compared to the β-grasp fold in ubiquitin domains (Figure 4.7C), the Flg_new and SHIRT domains share a similar β-strand topology but have lost the central helical region and are more elongated.

4.2.4 The MBG domain superfamily

The surface mucus adhesin Lar_0958 from *Lactobacillus reuteri* contains tandem domain repeats with a structure remotely related to Immunoglobulin-like β-sandwich domains, and some elements of similarity to β-grasp domains (Etzold

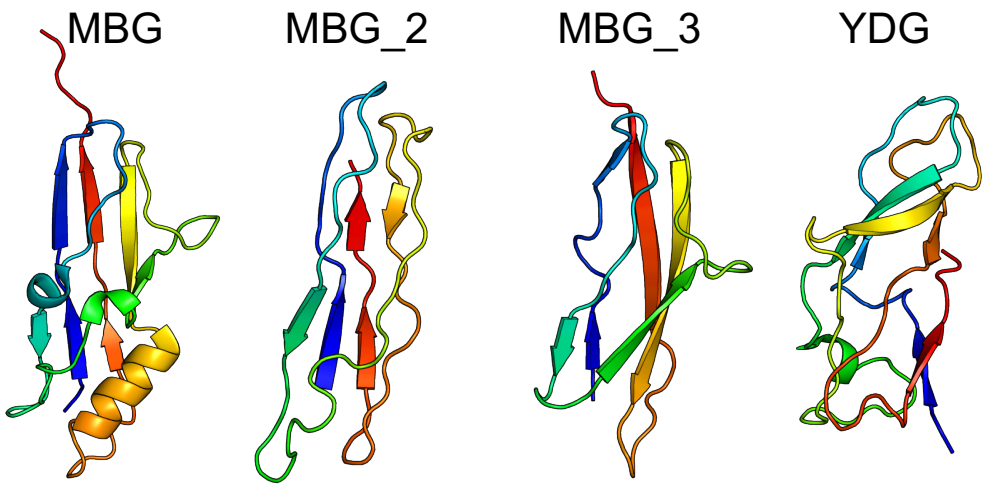

FIGURE 4.8 Experimental structure of the MBG domain and computational models for other families in the MBG clan. MBG domain (Pfam:PF17883) from the *L. reuteri* surface mucus adhesin Lar_0958 (PDB:4NG0), and structure predictions by trRosetta (J. Yang *et al.*, 2020) provided by the Baker lab for MBG_2 (Pfam:PF18676), MBG_3 (Pfam:PF18887), and YDG (Pfam:PF18657) domain families in Pfam. Domains share 20-30% sequence identity and structures are below 4Å C_α RMSD to each other, except for the YDG model, which cannot be aligned to any of the other domains, suggesting it is an incorrect fold prediction. Domain structures visualised using PyMol, rainbow colored from N-terminal (blue) to C-terminal (red) and shown in a similar orientation with N-terminus down and C-termini up.

et al., 2014). Alex and I have defined the domain as a novel fold, named Mirror Beta-Grasp (MBG), because the order of β-strands in the central β-sheet (B-A-E-D-C) is a partial mirror image of the β-grasp fold (C-D-A-B), and they share a central α-helix opposite to the β-sheet (Figure 4.8). Compared to Ig-like folds, the MBG domain is missing the D and E strands but conserves the A-G-F β-sheet, and has a helical insertion between the F and G strands.

Based on the Lar_0958 domain repeat structure, Alex built a new MBG family in Pfam (Pfam:PF17883), which was commonly found as tandem repeats. Alex and I further detected sequence similarities (in particular a highly conserved YDG motif) between the MBG family and others built from tandem repeats: MBG_2 (Pfam:PF18676), MBG_3 (Pfam:PF18887), and YDG (Pfam:PF18657). We classified these families in a new Pfam clan (Pfam:CL0682) and I compared their evolutionary relationships (Figure 4.9). The MBG_2 and MBG_3 domains are

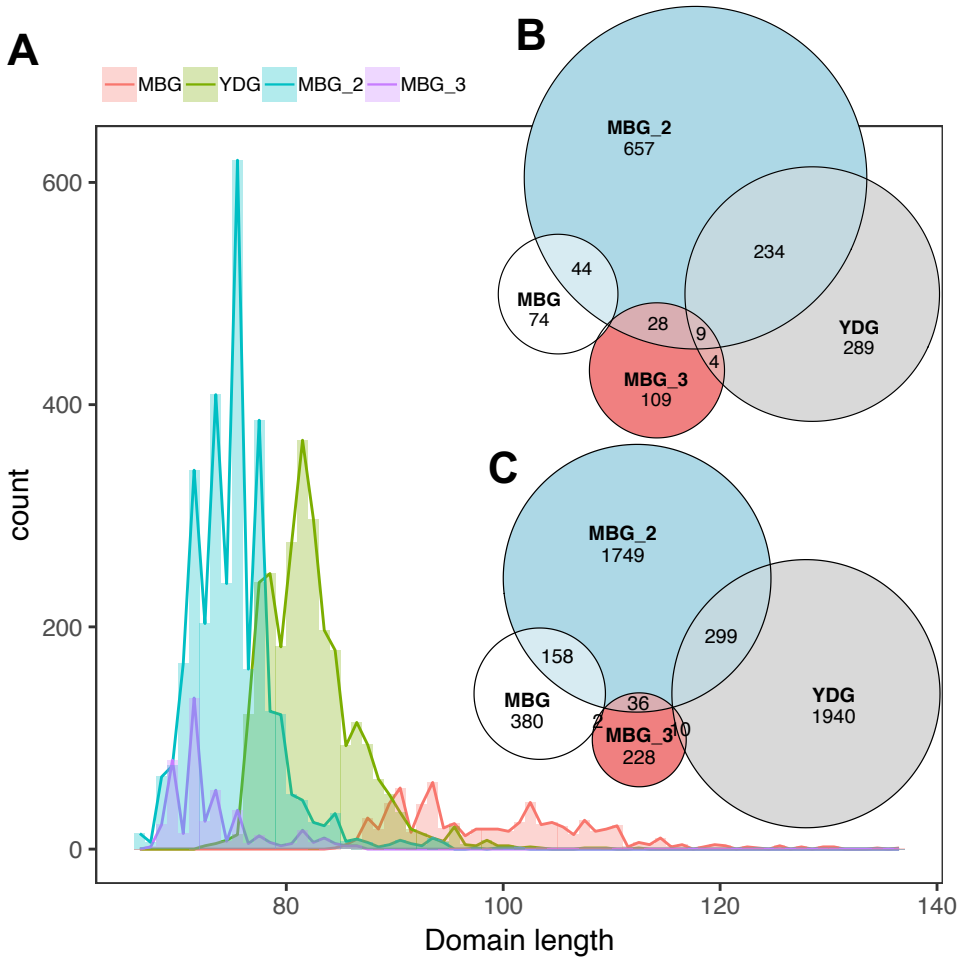

FIGURE 4.9 Comparison of the four Pfam families in the MBG clan. A) Distribution of domain lengths split by Pfam familiy. Venn diagrams of B) the number of adjacent domain pairs for each Pfam family (pairs of the same and different family), and C) number of domain co-occurrence in the same protein.

on average shorter (below 80 residues), and MBG domains are the longest, spanning a wide range of lengths above 100 residues (Figure 4.9A). The MBG_2 and YDG domains are the most widespread (over 1,300 domains from each family in UniProt), and MBG and MBG_3 domains are less common (around 500 domains from each family in UniProt).

Domains in the MBG_2 family are therefore likely ancestral, and domains in the three other families may have evolved independently from MBG_2 domains. MBG_2 domains are commonly found adjacent to domains of the MBG, MBG_3 and YDG families (Figure 4.9B), and as part of the same proteins (Figure 4.9C); and there is little overlap between the other families.

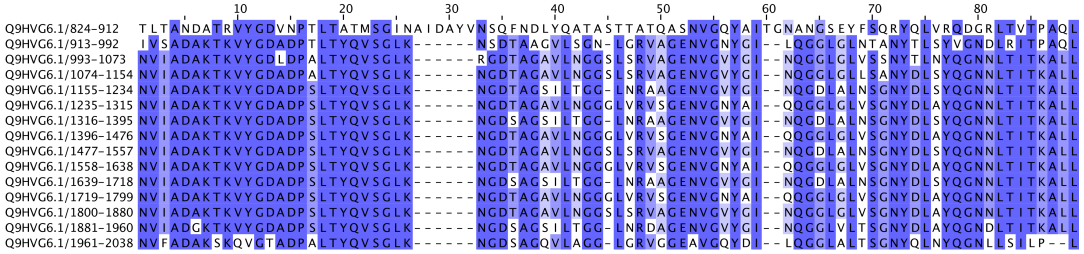

FIGURE 4.10 Alignment of tandem MBG_2 domain repeats in the surface protein CdrA from *Pseudomonas aeruginosa*. Alignment visualised using Jalview (Waterhouse *et al.*, 2009) and colored by sequence identity.

The predicted models of MBG_2, MBG_3 and YDG domains provided by the Baker lab using trRosetta (J. Yang *et al.*, 2020) suggest that these domains do not contain the helical region observed in the experimental structure of the MBG domain (Figure 4.8). Except for YDG, models appear to be reliable and predict β-sandwich folds closely related to Immunoglobulin domains. The helical insertion in the MBG domain would therefore represent a structural elaboration to the MBG_2 fold, predicted to be a shorter Immunoglobulin missing the D and E strands, similar to the case of the Rib domain.

The research lab of Tanmay Bahrat at the University of Oxford has been studying the *P. aeruginosa* surface protein CdrA, which contains highly similar MBG_2 tandem domain repeats, and recently imaged the protein forming interacting fibrils at the bacterial surface by Cryo-Electron Tomography (Melia *et al.*, 2021). They are now interested to experimentally determine a high-resolution structure of the different domains in the CdrA protein, including the MBG_2 domain repeats (Figure 4.10).

Based on our previous studies of domains in the MBG superfamily, I advised them on possible domain constructs to test experimentally (Table 4.1). I recommended to solve the first MBG_2 domain (R1), with a lower sequence similarity to the other repeats, the third domain (R3), and an additional tandem domain pair to investigate the domain orientation (R3-4). I added 4 N-terminal and 7 C-terminal extra residues to the Pfam domain boundaries, predicted to be too short based on sequence alignments to other MBG domains and later confirmed by trRosetta models. Constructs are currently being tested for expression and experimental structure determination by X-Ray crystallography in Tanmay Bahrat's lab.

TABLE 4.1 Sequence constructs for MBG_2 domain repeats for structure determination.

```
>CdrA_R1
QTATLTANDATRVYGDVNPTLTATMSGINAIDAYVNSQFNDLYQATASTTATQASNVG
QYAITGNANGSEYFSQRYQLVRQDGRLTVTPAQLI
>CdrA_R3
TPAQLNVIADAKTKVYGDLDPALTYQVSGLKRGDTAGAVLNGGSLSRVAGENVGVYGI
NQGGLGLVSSNYTLNYQGNNLTITKALLN
>CdrA_R3-4
TPAQLNVIADAKTKVYGDLDPALTYQVSGLKRGDTAGAVLNGGSLSRVAGENVGVYGI
NQGGLGLVSSNYTLNYQGNNLTITKALLNVIADAKTKVYGDADPALTYQVSGLKNGDT
AGAVLNGGSLSRVAGENVGVYGINQGGLGLLSANYDLSYQGNNLTITKALLN
```

4.3 Structure determination pipeline

Over the course of this project, Alex and I were compiling a list of interesting tandem domain repeats from our large-scale database searches and classification efforts. We were keen on studying these domains further, and in 2019 I started a collaboration with Matthew Bowler, a beamline scientist at EMBL Grenoble interested in testing a sequence to structure determination pipeline at the ESRF Synchrotron. I selected a subset of ten target domains and sent their sequences to Matthew, who took care of ordering constructs, expressing and purifying proteins at EMBL Heidelberg facilities, running crystallisation trials, and collecting X-Ray diffraction data at EMBL Grenoble.

Our collaboration with Matthew Bowler had two main goals. For us, it was an opportunity to increase the structural knowledge of domains that form repetitive stalks in bacterial surface proteins, specially focusing on domains lacking known homology to other families and cases of potential structural evolution, like previously observed in Rib domains. For Matthew, it was an opportunity to test the feasibility and costs of a "sequence to structure" service for future users of the ESRF Synchrotron.

TABLE 4.2 Target domains selected for structure determination.

	Construct ID	UniProt ID	Pfam	Organism	Length
1	ssspr51	V6Z1A5	SSSPR-51	Streptococcus agalactiae	56
2	ssspr51_2	V6Z1A5	SSSPR-51	Streptococcus agalactiae	104
3	ssure_spneu	Q8DRK2	SSURE	Streptococcus pneumoniae	154
4	ssure_soral	A0A0F2E6N5	SSURE	Streptococcus oralis	151
5	mbg2_cdra	Q9HVG6	MBG_2	Pseudomonas aeruginosa	81
6	ydg_bias	J3AS56	YDG	Caulobacter sp.	89
7	mbg3_short	A0A1I7BCE6	MBG_3	Algoriphagus locisalis	73
8	big6_cp_sasy	D2JAN8	Big_6	Staphylococcus aureus	90
9	rib_c1_saureus	A0A033U9S4	Rib	Staphylococcus aureus	87
10	sh3_domswap	Q08509	SH3_1	Mus musculus	122

4.3.1 Selection of domain targets

I selected a total of ten domains from families with unknown structure and/or homology to other families, structural variations of domains with known structure, such as predicted domain atrophy examples, and domains with extremely biased amino acid compositions, among others. The list of ten target domains and their sequence constructs are shown in Tables 4.2 and 4.3, respectively.

Domains include an isolated and a pair of tandem SSSPR-51 domains (Pfam:PF18877), predicted as domain atrophied MucBP domains (Pfam:PF06458) and found in tandem identical repeats; two SSURE domain repeats (Pfam:PF11966) from the PavB Periscope protein with unknown structure and homology to other families (Bumbaca *et al.*, 2005); and three domains from the MBG clan including an MBG_2 domain repeat from the CdrA protein, a domain repeat with biased sequence from the YDG family, and a short MBG_3 domain. I additionally selected a circularly permuted Big_6 domain repeat (Pfam:PF17936) from the SasY Periscope protein, a Rib domain from the second SSN cluster corresponding to *Staphylococcus aureus* surface proteins, and an artificial tandem pair of identical SH3 domains from the human EPS8 protein (PDB:1I07) predicted to form a domain swap, which would be the first structure of a tandem domain swap in the PDB, if successful.

TABLE 4.3 **Sequence constructs for target domains selected for structure determination.**

```
>ssspr51
PAKKVVTNHVDEDGNPIAPQEEGTTPNKSIPGYEFTGKTVTDPDGNTTHIYRKVKK
>ssspr51_2
PAKKVVTNHVDEDGNPIAPQEEGTTPNKSIPGYEFTGKTVTDPDGNTTHIYRKVKKV
VTNHVDEDGNPIAPQEEGTTPNKSIPGYEFTGKTITDKDGNTTHIYRKINN
>ssure_spneu
LISKETVQKAVADNVKDSIDVPAAYLEKAKGEGPFTAGVNHVIPYELFAGDGMLTRL
LLKASDKAPWSDNGDAKNPALSPLGENVKTKGQYFYQVALDGNVAGKEKQALIDQFR
ANGTQTYSATVNVYGNKDGKPDLDNIVATKKVTININGLI
>ssure_soral
LTTPAEVQKGVADNTKDTVDVPASYLDKANFPGPFTAGVNQVIPYEFFAGDGMLTRL
ILKASDKAPWSDNGSAKNPALPPVEKLGKGLYFYEVDLAGTQGKSDKELLDLLKQNG
TQSYKATIKVYGAKDGKPDLTNLVATKDLTVNLNGLT
>mbg2_cdra
LNVIADAKTKVYGDLDPALTYQVSGLKRGDTAGAVLNGGSLSRVAGENVGVYGINQG
GLGLVSSNYTLNYQGNNLTITKAL
>ydg_bias
LTAGLTGTVTKTYDGTTVATLGNNLGLSGLVAGDTVSVASTGAAYADKNAGAGKTVT
ASGVNLGGADAGNYVLASTTASAAVGQINAKT
>mbg3_short
LEGITFADAVFVFDGTAKSLAIGGTIPEGTSVSYANNSRTNVGTQEVTATISGSNFT
TLVLTADLTITPATIT
>big6_cp_sasy
KGNWTVDVPEGTELKVGNEITATETDMSGNKSESGKGKVTDTTAPEAPSVNDTEVGS
KKVSGKGHEVGNTVTVTFPDGKTATSKVDEKGN
>rib_c1_saureus
MENSQYEPTTDGVTKDHGTPTTSDDVTGSVTIPDYPTDKDQPTITVDDETQLPDGNT
PGTTEVDVTVTYPDGTQDHIKVPVTVGEQA
>sh3_domswap
KKYAKSKYDFVARNSSELSVMKDDVLEILDDRRQWWKVRNASGDSGFVPNNILDIMR
TPEGGKKYAKSKYDFVARNSSELSVMKDDVLEILDDRRQWWKVRNASGDSGFVPNNI
LDIMRTPE
```

4.3.2 Progress and results

Initially, Matthew estimated the time to order constructs (synthetic genes) to take up to 2 weeks, the protein expression and purification at EMBL Heidelberg to be up to 2 months, and the protein crystallisation trials (the most uncertain step) to be around 2 months. Shooting the protein crystals with X-Rays at the ESRF Synchrotron would depend on time availability, but was estimated to be relatively straightforward using robots that handle the crystallization plates automatically at EMBL Grenoble facilities. In total, he estimated that the pipeline for structure determination would take at least six months.

I sent the list of ten domain sequences to Matthew in August 2019, and he introduced removable Histidine tags at the N-terminus to facilitate purification using NiNTA (nickel-charged) affinity chromatography. In October 2019, the constructs had been ordered and tested for expression at small scale, with promising results on all but the biased YDG domain (Figure 4.11). Three months later (January 2020), seven out of the ten domains had been successfully expressed in larger quantities and concentrated (the MBG_2 domain of CdrA and Big_6 domain of SasY were not successful), and were ready to start crystallisation trials. On the 3rd of March 2020, Matthew told us that three of the domains had crystallised successfully (the short MBG_3 and two SSURE domains) and were ready to collect X-Ray diffraction patterns at the Synchrotron. Even though the crystals are small and difficult to see by eye (Figure 4.12), Matthew was hopeful they would diffract and that even if they did not, it would be possible to optimise the crystallisation conditions to grow bigger crystals.

Unfortunately, the same week Matthew were planning to obtain diffraction data of the crystals, the EMBL Grenoble shut down due to the coronavirus pandemic, and the progress of the project was paused temporarily. Even though the ESRF Synchrotron partially reopened its activity towards the end of 2020, projects related to coronavirus research were prioritised and it was not possible to run our experiments. Overall, we have successfully expressed and purified seven out of the ten domain constructs, and obtained three crystals (Table 4.4); this success rate is higher than expected given that some constructs were of high risk. We are hopeful that we can resume the project in the near future and obtain high-resolution experimental structures for some of these domain targets.

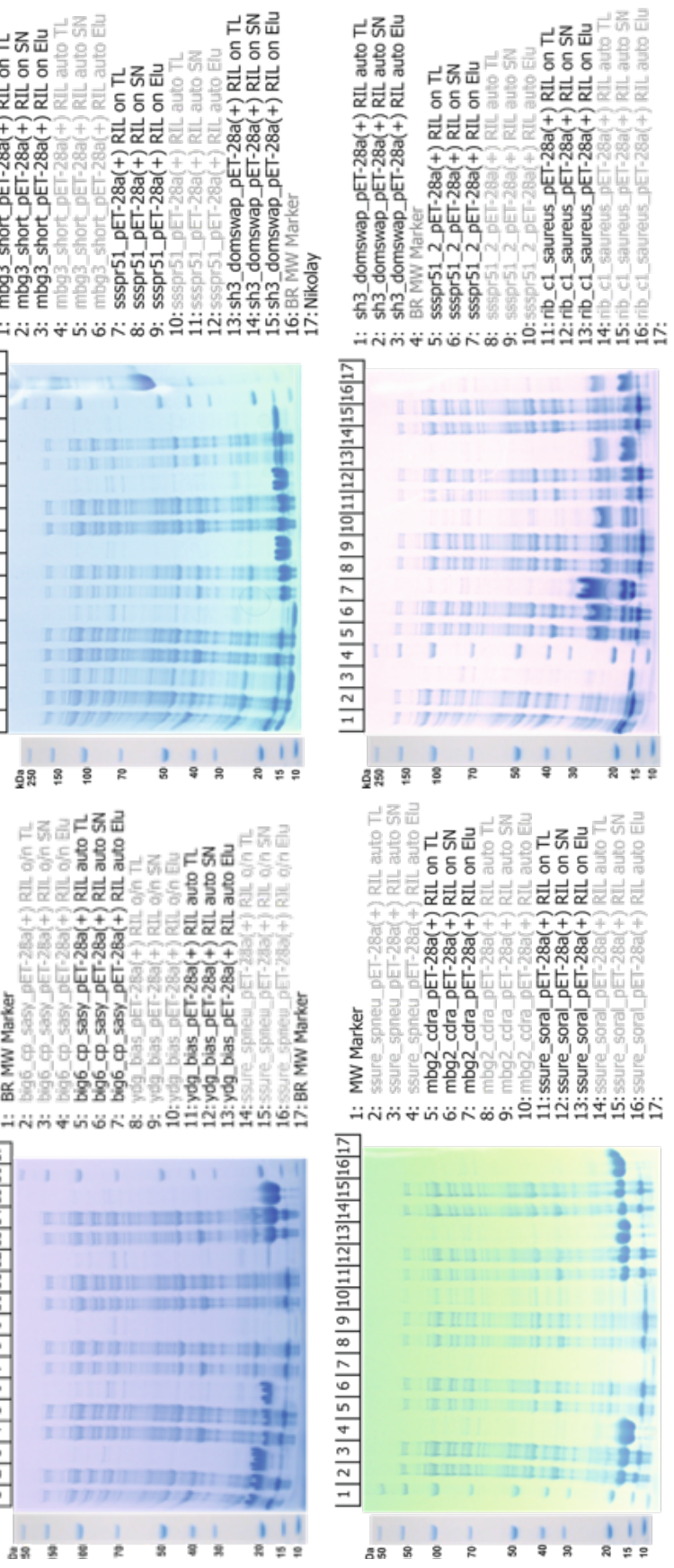

FIGURE 4.11 **Results of the test expression experiments at the EMBL Heidelberg PepCore facility.** SDS-PAGE gels for total lysate (TL), i.e. all of the broken bacteria; supernatant (Sn), i.e. solution left after breaking cells and centrifuging; and eluted solution after purification using NiNTA affinity chromatography (Elu). Pictures taken by Matthew Bowler and shared through our scientific collaboration.

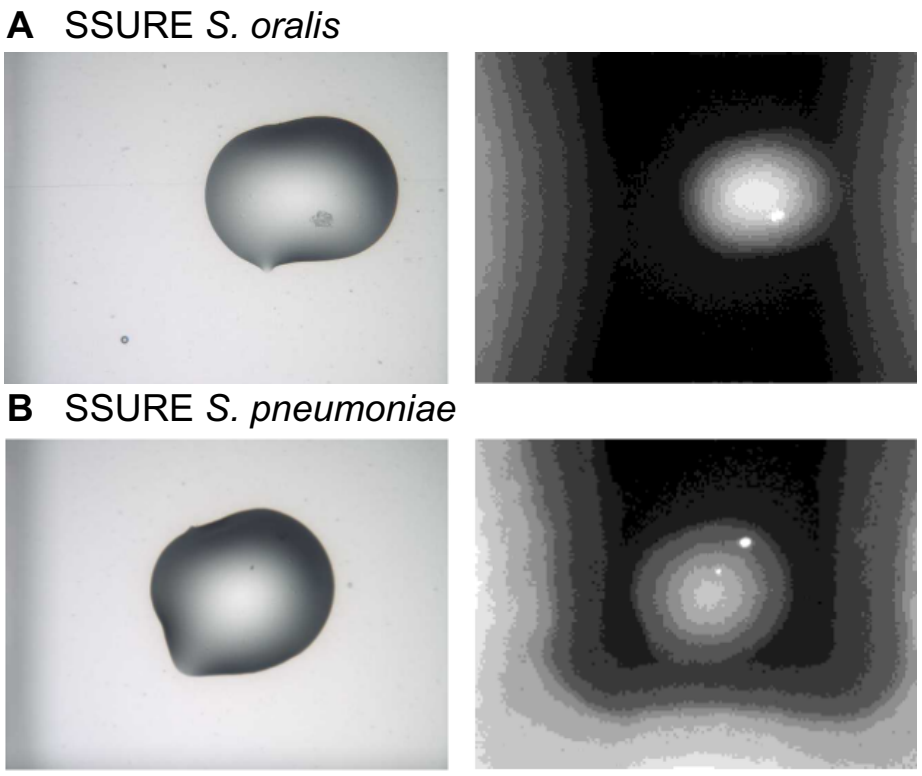

FIGURE 4.12 Images of crystals for the two SSURE domain targets in the structure determination pipeline. Bright spots on the dark images to the right correspond to small protein crystals. Pictures taken by Matthew Bowler and shared through our scientific collaboration.

TABLE 4.4 Status of the structure determination pipeline for each domain target at the time of writing this thesis. HTX: High Throughput Crystallisation trials. Yes (Y), No (N), in progress (?).

	Construct ID	Cloned	Test expression	Expression	HTX	Crystals	Model
1	ssspr51	Y	Y	Y	Y	?	?
2	ssspr51_2	Y	Y	Y	Y	?	?
3	ssure_spneu	Y	Y	Y	Y	Y	?
4	ssure_soral	Y	Y	Y	Y	Y	?
5	mbg2_cdra	Y	Y	N	-	-	-
6	ydg_bias	Y	N	-	-	-	-
7	mbg3_short	Y	Y	Y	Y	Y	?
8	big6_cp_sasy	Y	Y	N	-	-	-
9	rib_c1_saureus	Y	Y	Y	Y	?	?
10	sh3_domswap	Y	Y	Y	Y	?	?

4.4 Discussion

Experimental structures of tandem domain repeats have shown that long repeats in bacterial surface proteins fold into stable globular domains, providing insights into their role as rigid stalks; they further revealed surprising cases of structural malleability and domain evolution. In this chapter, I have analysed in detail several structures of domains that form tandem repeats, some of them experimentally solved by our collaborators and others available in the PDB. I have further attempted to experimentally solve the structures of interesting tandem domain repeat families through collaborations with structural biology groups.

Tandem domain repeats show remarkable examples of structural malleability. The evolutionary recent cases of domain atrophy in the SasG_E and Rib domains are very rare, as described by Prakash and Bateman (2015), and might be the consequence of unusual evolutionary and selective pressures in proteins with highly similar tandem domain repeats. Other domains, such as SHIRT and domains in the MBG clan, further show atypical folds only remotely related to widespread topologies such as β-grasp ubiquitins and β-sandwich Immunoglobulins, but with little structural similarity to other domains in the PDB.

Several indicators have shown an extraordinary high stability of tandem domain repeats, such as Rib and SHIRT, which could be related to their structural malleability and ability to tolerate large deletions and insertions. In particular, it was remarkable that the SHIRT domain remained folded and crystallised with an artificial deletion of its entire N-terminal strand, and surprising that its structure could be solved at all with wrong domain boundaries. There are important open questions about what the most stable and malleable parts of these domains are.

Due to unforeseen circumstances, our structural determination pipeline could not be finalised in time for this thesis. I expected some of the target domain structures to provide additional data to understand the structural malleability and flexibility of tandem domain repeats. However, I was encouraged by the high success rate of protein expression and purification (9/10 in test expression and 7/10 in purification), which further indicated that domains I selected are likely folded into stable globular structures, despite the challenges in selecting correct domain boundaries. Protein crystals were obtained for SSURE domains, an interesting long repeat with fibronectin binding function (Bumbaca *et al.*, 2005), which has very few homologous sequences and that are believed to be a double domain.

The number of tandem domain repeat structures solved experimentally is still small, and some of the domains are difficult to model and solve experimentally due to their limited homology and large structural changes to other known domain families and structures. Structure prediction techniques have become more accurate and offer the potential to revolutionise the types of computational analyses presented in this chapter, shifting studies of protein function and evolution from the sequence perspective to the structural and molecular details. I have presented some examples of this trend in the analysis of MBG domains, where models by trRosetta provided by the Baker lab have been key to interpret the structural differences among the domain families. Some families, however, could not be modelled due to the limited number of sequences, such as the SSURE domains, and others proved difficult to model due to their atypical sequence composition and large structural changes, such as the likely wrong model of the YDG family. In addition, the number of experimental structures for constructs of tandem domains is still very limited, hindering important analyses of inter-domain linkers, domain interactions and orientations.

Results presented in this chapter should motivate further experimental and computational efforts to characterise the structures of tandem domain repeats. I have started some of these efforts through our collaborators, but expect other research groups to become interested in these domains and attempt to discover further cases of structural malleability and evolution, and study their implications for protein stability and function.

4.5 Methods

4.5.1 Structure alignment and similarity

Protein structures were visualised using PyMol version 2 (downloaded from https://pymol.org/2, and structural alignments were built using PyMol's built-in CE-align tool (Shindyalov and Bourne, 1998). Sequence alignments were visualised in Jalview (Waterhouse *et al.*, 2009).

The root-mean-square deviation (RMSD) based on C_α atoms was used to evaluate structural similarity between protein structures. The RMSD is the measure of the average distance between the atoms of a pair of superimposed proteins. For a pair of vectors of point coordinates x and y, the RMSD depends on the

number of aligned points n and their distances and can be calculated as following:

$$RMSD(x, y) = \sqrt{\frac{1}{n}\sum_{n}^{i=1} \|x_i - y_i\|^2} \tag{4.1}$$

4.5.2 Sequence similarity network

The sequence similarity network of Rib domains (Figure 4.5) was constructed using the domain sequences from the three Rib domain families in Pfam: Rib (Pfam:PF0428), RibLong (Pfam:PF18957) and aRib (Pfam:PF18938). A subset of 400 sequences from each family were randomly selected and used to run an all-against-all BLASTp search (Altschul *et al.*, 1997). Sequence similarity hits with a score above 35 bits were further used to build a network with the `igraph` package in `R`, and displayed using the default Perfuse Force Directed Layout method. Nodes of the network were further colored according to their Pfam family.

Chapter 5

Determinants of misfolding in tandem domains

> *"Folding is the final stage in the translation of genetic information to a working protein and is one of the simplest examples of biological self-organization."*
>
> - Bryngelson and Wolynes (1987): "Spin glasses and the statistical mechanics of protein folding"

In this chapter, I investigate protein misfolding determinants in tandem domains. I present a new computational method, developed in collaboration with Robert Best and Pengfei Tian at the National Institutes of Health (NIH, USA) and named TADOSS (TAndem DOmain Swap Stability predictor), to estimate the propensity of globular domains to form tandem domain swaps, a type of misfolded protein conformation. I use TADOSS to identify misfolding-resistant domains across protein families, with special focus on tandem domain repeats, and describe potential determinants of protein misfolding.

Contents in this chapter have appeared in two separate scientific publications: a short two-page Applications Note in *Bioinformatics* describing the TADOSS method (Lafita, Tian, Best, and Bateman, 2018), and a review article in *Current Opinion in Structural Biology* about tandem domain swapping (Lafita, Tian, Best, and Bateman, 2019). I developed and wrote the code of the TADOSS method based on scripts for calculating alchemical free energies provided by Pengfei Tian and Robert Best, and I wrote the manuscripts and generated the figures with input from co-authors.

5.1 Introduction

Protein folding is a complex physical process, commonly defined as traversing a rugged multidimensional energy landscape with potential local minima (Bryngelson and Wolynes, 1987). When trapped in these local energy minima, proteins adopt meta-stable conformations other than their functional or native structure, regarded as misfolded or non-native states. In addition to being nonfunctional, misfolded proteins can lead to other detrimental effects for cells and organisms, such as protein aggregation (Bennett *et al.*, 2006; Dobson *et al.*, 2019).

Modifications to the sequence of a protein, even if small, can have large effects on a protein's energy landscape and impact its folding and stability in unforeseeable ways. Protein misfolding is therefore a longstanding challenge in the engineering of proteins and the treatment of protein-related diseases. Understanding the structural and energetic determinants of protein folding, and thereby misfolding, is hence of broad biomedical and biotechnological interest.

Although nearly half of all proteomes are composed of multidomain proteins (an even higher fraction in Eukayotes), domains are generally regarded as independently folding units and our knowledge of protein folding is based primarily on studies of isolated domains in solution. However, several studies report interactions between adjacent domains in multidomain proteins, some of which occur during protein folding and increase the likelihood of forming misfolded conformations (Han *et al.*, 2007; Batey *et al.*, 2008). Cellular mechanisms have also been described to play an important role in the correct folding of multidomain proteins, such as co-translational folding on the ribosome (Waudby *et al.*, 2019).

In multidomain proteins, the protein misfolding rate has been experimentally shown to be higher for proteins with highly similar adjacent domains. Wright *et al.* (2005) used aggregation kinetics experiments of a tandem pair of homologous immunoglobulin domains (titin I27) to show that the sequence similarity between adjacent homologous domains is a strong determinant of protein misfolding. Further studies with similar I27 protein constructs by M. B. Borgia *et al.* (2011) and A. Borgia *et al.* (2015) identified specific misfolded conformations using single-molecule fluorescence. These misfolded conformations were found to be caused by native-like contacts between adjacent domains, and remained stable for longer times in pairs of identical domains.

The importance of the sequence identity of adjacent domains for protein mis-

folding reveals a fundamental attribute of their folding mechanism, the formation of so-called native interactions: energetically favourable contacts between specific residues of the domain. When two identical domains are spatially close, such as adjacent in a multidomain protein, these native contacts can be satisfied within domains (intra-domain) or across domains (inter-domain), with little disruption to their environment. As a result, native domain interactions can also be satisfied by misfolded non-native conformations where domains are folded intertwined into each other.

Two plausible mechanisms by which similar sequences might interact are via the formation of parallel, in-register β-sheet structures (amyloid-like), or by the reciprocal exchange of equivalent secondary structure elements, forming what is known as a domain swap (Figure 5.1). Even though these conformations are generally less stable than the native two-domain protein, they correspond to local minima in the folding energy landscape far from the native conformation and trap proteins for significant periods of time.

Tandem domain swaps are one of the best characterised misfolded conformations in tandem domains. They are transiently stable and suitable for theoretical and computational studies thanks to the high fraction of conserved native contacts, and serve as an idealised model for other misfolded states with a lower fraction of native contacts, such as misfolded intermediate with a single folded domain shown in Figure 5.1. Their formation, or infeasibility, is therefore a strong determinant of misfolding in proteins with tandem homologous domains.

In this chapter I focus on understanding tandem domain swaps theoretically to gain insights into the misfolding propensity of tandem domain repeats. I explore how tandem domain swaps are formed and how stable they are across different protein domains, with the aim to find common determinants of misfolding in proteins.

In the following introductory subsections, I describe tandem domain swapping and its relation to two other well-studied protein conformational variants: domain swap oligomers and circular permutations. I review experimental techniques and evidence for the formation of tandem domain swaps *in vitro* and their link to protein misfolding and aggregation, and describe various approaches from the literature to computationally model and predict the formation of tandem domain swaps in protein domains.

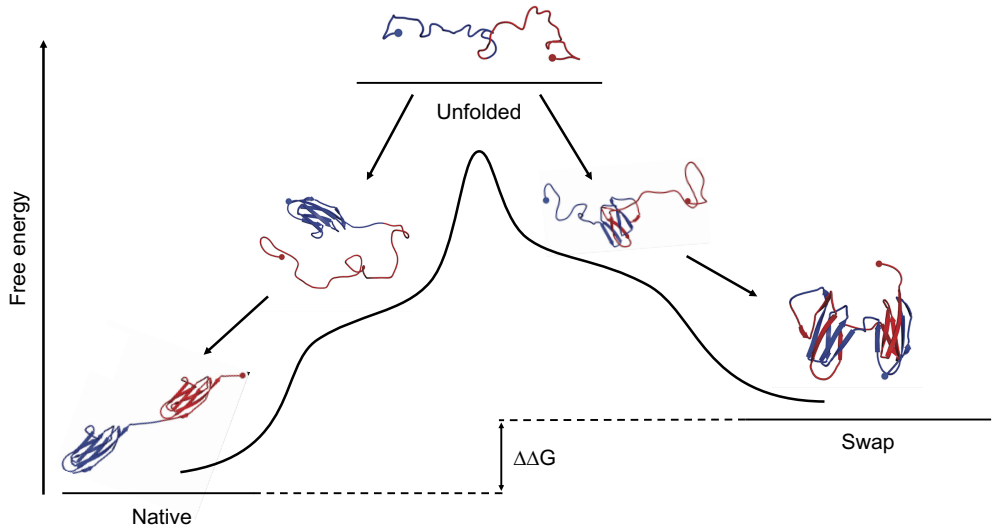

FIGURE 5.1 Schematic representation of the folding energy landscape of tandem domain swapping in a pair of adjacent identical protein domains. The unfolded protein (top) can fold into the "native" conformation with two sequential domains (bottom left), or into a meta-stable misfolded "swap" conformation with two intertwined domains (bottom right). Intermediate states for each conformation correspond to folding a single domain. The stability difference between the native and swap states is shown as a free energy difference ($\Delta\Delta G$). Structures correspond to a tandem pair of Immunoglobulin-like domains of the human titin protein. Figure adapted from Lafita *et al.* (2019).

5.1.1 Tandem domain swapping

Domain swapping refers to the reciprocal exchange of secondary structure elements between protein domains, forming an intertwined conformation. Although there does not seem to be a unified mechanism for domain swapping (Mascarenhas and Gosavi, 2017), in some cases complete unfolding has been shown to be a requirement (Liu *et al.*, 2012). The swapping mechanism requires the extension of a short region of the native domain fold to form what is known as the "hinge loop", which bridges the two interacting domains (Figure 5.2a). The properties of the hinge loop, such as its length and sequence composition, are critical for the formation of domain swaps; for instance, Prolines in hinge loops have been found to favour domain swapping (Rousseau *et al.*, 2001). Attempts to engineer domain swaps through the introduction of designed hinge loops in protein domains have been largely successful (Ha *et al.*, 2015; Nandwani *et al.*, 2019). The core of a domain swap is formed by a high fraction of native inter-domain interactions, as opposed to all intra-domain contacts in the two-domain conformation. Since the majority of the native contacts are fulfilled in the domain swap, the overall 3D structure of the domain is preserved, except for the region around the hinge loop, which makes domain-swapped conformations amenable to computational modelling.

Domain swapping can also occur between domains in separate protein chains in solution, forming an oligomer (Figure 5.2b). Two or more chains can be involved in domain swapping, possibly forming intricate multimeric assemblies. A diverse set of protein domains have been found to oligomerise through domain swapping, including all-α, all-β and mixed α/β folds (Gronenborn, 2009), with several hundred experimental structures deposited in the Protein Data Bank (PDB) and collected in the 3DSwap database (Shameer *et al.*, 2011). Some domain swap oligomers correspond to the native protein conformation, while others are less stable and found only as crystallisation artifacts.

Domain swapping can also occur in multidomain proteins through interaction of adjacent domains, as shown in Figure 5.2a. The formation of tandem domain swaps requires an additional linker that joins the termini of the domain, which sits in the middle of the protein chain, named the "central domain". The other "terminal domain" is domain-swapped, but conserves the termini of the native fold.

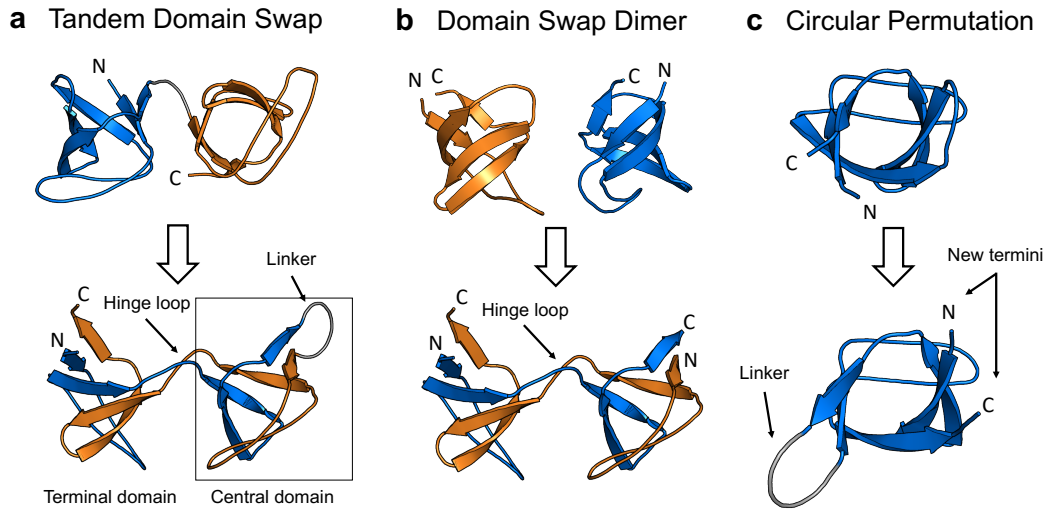

FIGURE 5.2 Comparison of domain swapping types and circular permutation. Native (top) and swapped or permuted (bottom) structural conformations of (a) tandem domain swaps, (b) domain-swapped dimers and (c) circular permutations. The central domain of a tandem domain swap is equivalent to a circular permutation, formed from the central region of the two-domain sequence comprising the C-terminus of the first domain and the N-terminus of the second domain, joined by the inter-domain linker. The terminal domain comprises the remaining parts of the sequence and folds with the termini in their native location. The hinge loop that connects the central and terminal domains is characteristic of domain swapping, and also present in domain-swapped oligomers. Structures shown are models of an SH3 domain (PDB:1SHG) generated in simulations by Tian and Best (2016) and visualised using PyMol. Figure adapted from Lafita *et al.* (2019).

The formation of the central domain and its terminal linker resembles a protein circular permutation (Figure 5.2c). Circular permutations are cyclic rearrangements of a domain sequence, so that the N and C-terminal regions are interchanged at a specific "cut" position, and that fold into the same structural unit with different domain termini (Bliven and Prli, 2012). Protein domains have been shown to be remarkably insensitive to circular permutations and retain their stability and activity in viable variants (Lo *et al.*, 2012a). Therefore, circular permutations occur frequently in natural proteins and have been successfully engineered in the lab (Uliel *et al.*, 2002), with thousands of experimental structures of circularly permuted variants across domain folds deposited in the PDB and collected in the CPDB database (Lo *et al.*, 2009). It is important to note that only a small subset of all the possible circular permutations along the sequence of a protein domain will be viable, meaning that variants can fold into stable domains, which has important implications for the formation of tandem domain swaps.

Although domain swapping has been extensively studied in protein oligomerisation, far less is known in the context of multidomain proteins and its connection to protein misfolding and aggregation. At the time of this thesis, there are no experimental structures of tandem domain swaps available in the PDB, unlike for circular permutations and domain-swapped oligomers, which further complicates their study. However, recent experiments have provided evidence for the formation of tandem domain swaps and their implication in the misfolding rate of multidomain proteins, providing a starting point for further investigation. I review some of the most important experimental studies in the next section.

5.1.2 Experimental evidence for domain swap misfolding

Tandem domain swapping events have been experimentally identified with cutting-edge single-molecule biophysical techniques. Evidence of misfolding events in tandem domains was first obtained using Atomic Force Microscopy (AFM) experiments on a pair of tandem identical immunoglobulin-like I27 domains from titin (Oberhauser *et al.*, 1999). In this work, a protein construct with the tandem domain pair was folded and refolded using a stretching force applied at the protein termini, measuring the resistance of the protein to the force at different timepoints. In several of the protein folding and refolding cycles, the force

corresponding to the unfolding of a single native domain led to an increase in molecular length corresponding to two domains, revealing the existence of a stable misfolded state, possibly native-like, involving interactions between the two adjacent domains. The nature of the misfolded state was however not clear at the time.

More recently, M. B. Borgia *et al.* (2011) and A. Borgia *et al.* (2015) have used single-molecule Förster Resonance Energy Transfer (FRET) experiments to examine the misfolding of the same pair of tandem I27 domains. Fluorescent FRET labels were attached near the termini of the protein and, upon refolding the protein from a chemically denatured state, two peaks at different FRET transfer efficiencies, which reflects the distance between FRET labels, were observed. The low efficiency peak corresponded to the native two-domain conformation, and was the most populated, while a second less-populated peak with high FRET efficiency was also observed. The FRET transfer efficiency of the second peak was found to be the same as when FRET labels were attached to equivalent residues in a single I27 domain. This finding and the longevity of the misfolded state were very suggestive of an intertwined conformation between the two domains involving domain swapping. These time-resolved FRET experiments also revealed additional misfolded states, attributed to amyloid-like species, although these were relatively less stable and for shorter times (A. Borgia *et al.*, 2015).

Immunoglobulin-like I27 domains from titin have so far been the prototype and only example used in experimental studies of tandem domain misfolding. However, recent experiments with similar single-molecule techniques suggest these findings are applicable to other protein domain folds, for example crystallins (Garcia-Manyes *et al.*, 2016). Although single-molecule experiments can infer the formation of domain-swapped protein conformations, they are unable to distinguish among the many possible swapped conformations, or determine which one is the most probable. Computational and theoretical techniques, described in the next section, have been used in combination to these experimental results to gain insights about the molecular mechanisms of protein misfolding.

5.1.3 Computational studies of domain swapping

Accurately modelling the structure and dynamics of proteins is still an unsolved problem in biology. The conformational space of proteins is too large and protein

folding rates are too slow to simulate in reasonable amounts of time, even in the most powerful modern computers (Brini *et al.*, 2020). Despite these challenges, approximate protein energy potentials, such as Rosetta (Alford *et al.*, 2017), allow relatively accurate protein simulations and stability calculations with realistic computing resources.

Molecular dynamics (MD) simulations are a computational technique used to study protein folding and conformational flexibility that consists in sampling the physical movements of atoms and molecules by integrating approximate energy potentials (also known as force fields) over time (Abraham *et al.*, 2015). They have been widely applied to predict the viability of circular permutations (Hu *et al.*, 2007) and to understand the molecular mechanisms of domain swapping (S. Yang *et al.*, 2004; Malevanets *et al.*, 2008). When applied to tandem domain repeats, Zheng *et al.* (2013) theoretically demonstrated that adjacent identical domains have frustrated folding energy landscapes, meaning that they contain additional local minima in the free energy landscape that can trap the protein into meta-stable misfolded states predicted to be "amyloid-like" and domain-swapped conformations.

Tian and Best (2016) later used simulations with coarse-grained energy functions based on the native domain structure, so-called "Gō models", to study the formation of tandem domain swaps and gain insights into experimental observations by A. Borgia *et al.* (2015). In the same study, Tian and Best (2016) went a step further and used their Gō model simulation analysis to compare the propensity of tandem domain swap formation across a set of seven representative domain folds, including Immunoglobulin-like, Ubiquitin-like, SH2, SH3 and PDZ superfamilies. Interestingly, they found that domains naturally forming tandem repeats show a higher misfolding resistance to domain swapping.

Although MD simulations have shed light into the molecular mechanisms of tandem domain swap formation, they are computationally expensive and require sophisticated software pipelines with significant manual intervention, limiting their scalability and applicability to a broader diversity of domains. Insights from MD simulations can serve as the basis to develop computational methods for the systematic prediction of tandem domain swaps from structural features alone, a more scalable and universal approach. Such structure-based prediction methods have been developed for circular permutations and domain-swapped oligomeri-

sation prediction, and include methods by Paszkiewicz *et al.* (2006) and Lo *et al.* (2012b) (CPred) to predict viable circular permutation "cut" positions along a domain sequence, and methods by Ding *et al.* (2006) and Baiesi *et al.* (2016) to predict the feasibility of forming "hinge loops" in domain swaps.

These prediction methods use graph theoretical approaches on the residue interaction network calculated from the native domain structure, with some other additional features to improve the performance. The results are presented as sequence profiles, where the propensity to form a viable circular permutant or domain swap is assigned to each residue along the domain sequence, useful to identify both the number and location of stable swaps and permutants. Although these methods could potentially be useful at the task of predicting tandem domain swaps, which combine features of both circular permutations and domain swap oligomers, only the CPred method is openly available as a web-server.

As part of their comprehensive simulation study of tandem domain swapping, Tian and Best (2016) described an "alchemical" model — a simplified energetic calculation that does not require simulation — to estimate the free energy difference between the native and domain-swapped conformations. They found that their "alchemical" free energy correlated well with their simulations and that it could be more generally used to predict tandem domain swap misfolding. Pengfei Tian and Robert Best had applied their "alchemical" model to the seven domain folds in their simulation study, but used pre-defined manual "hinge-loop" positions based on their knowledge, and it was unknown if the approach could be generally applicable to other domains. I contacted them to express our interest in using their model to do large-scale analyses of tandem domain swapping in protein domains, and we decided to start a collaboration to implement an automated version of the method, which is described in the next section.

5.2 Prediction of tandem domain swapping

The "alchemical" free energy approach described by Tian and Best (2016) consists in using an artificial (non-physical) pathway of structural changes, instead of considering the physical but intractable pathway of unfolding the native domains and refolding them in a domain-swapped conformation, avoiding the use of molecular simulations. The alchemical pathway proposed by Tian and Best

(2016) consists in two steps: the formation of a new termini loop from the inter-domain linker (join) and the unfolding of a short segment of the native domain to form a hinge loop (cut). The total free energy difference between the native and swapped conformations ($\Delta\Delta G$) can be then split into the energy difference of the join (ΔG_J) and cut (ΔG_C) operations.

Pengfei Tian and Robert Best had implemented a script to calculate the alchemical $\Delta\Delta G$ for a few specified cut positions in each of the seven domains in their study (Tian and Best, 2016). With their help, I developed a separate standalone method, named TAndem DOmain Swap Stability predictor (TADOSS), using their original code implementation. TADOSS incorporates a few key advancements to automate the calculations on any query domain and to improve the analysis and visualisation of predictions, which I describe in detail in the next subsections.

5.2.1 The TADOSS method

The aim of TADOSS is to estimate the relative free energy difference between the native two-domain and possible domain-swapped conformations of a domain. First, a matrix of pairwise residue contact energies is estimated using a coarse-grained Gō model from the input structure of a single domain, as illustrated in Figure 5.3 and described in the Methods section. TADOSS then systematically evaluates all possible "cut" positions along the sequence of the domain from the matrix of contact energies, calculating the free energy difference ΔG_C upon formation of a hinge loop of at least three residues centered at the position. The ΔG_C profile is valuable to identify the regions of the domain that are more susceptible to form hinge loops and hence the most probable tandem domain swap conformations (Figure 5.3). Finally, TADOSS estimates the free energy difference of joining the termini of the central domain ΔG_J and calculates the total free energy difference $\Delta\Delta G$ of forming tandem domain swaps by summing up the free energy differences from the cut and join steps. The most probable domain swaps are those with the highest $\Delta\Delta G$ (most positive). More details about the alchemical free energy model and its thermodynamic parameters can be found in the Appendix C.

The alchemical free energy difference from TADOSS correlates well with the free energies obtained in MD simulations, as originally shown by Tian and

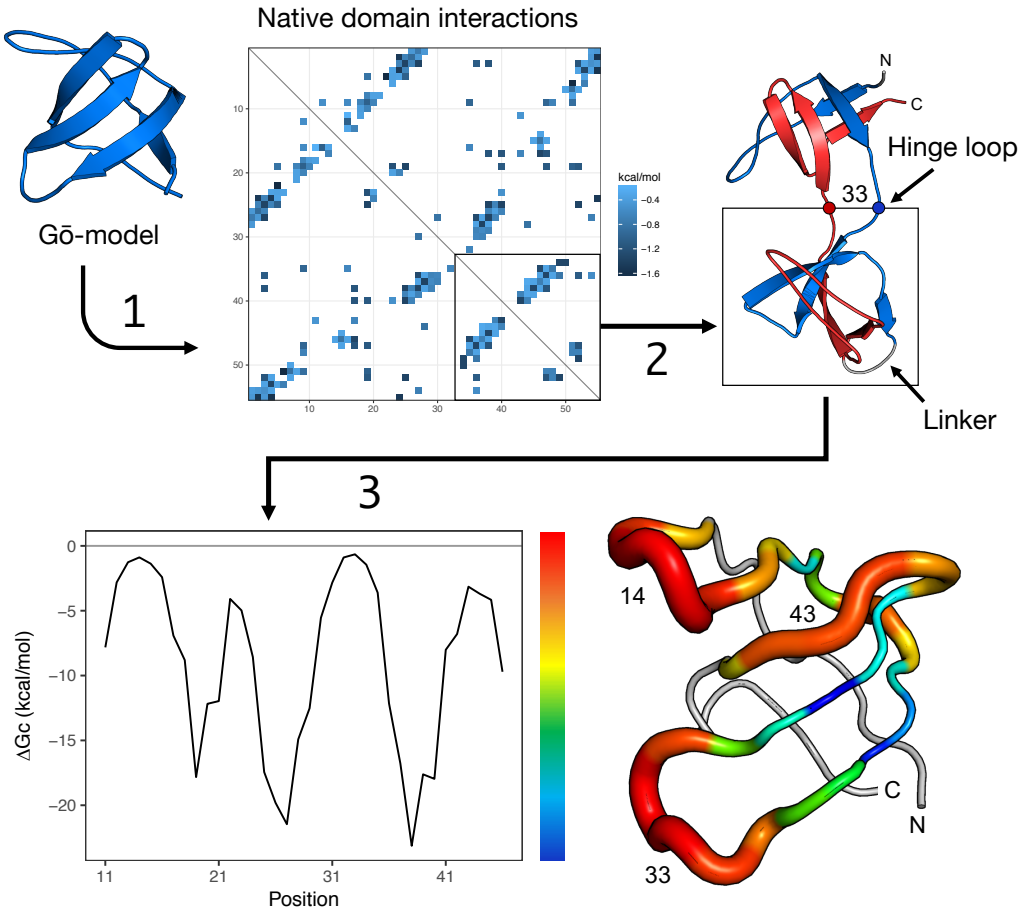

FIGURE 5.3 Steps of the alchemical free energy estimation by TADOSS. 1) The energy of the native residue contacts in the domain is calculated using a Gō model from the structure of a single protein domain, 2) the energy contributions of forming a hinge loop and connecting the domain termini with a linker are estimated from the distorted native contacts, and 3) a free energy profile along the domain sequence is constructed by systematically evaluating all possible hinge loop positions, where higher free energy differences correspond to more stable hinge loops. The energy profile can be mapped in 3D to the structure of the domain to visually identify hinge loop hotspots using PyMol. Figure adapted from Lafita *et al.* (2019).

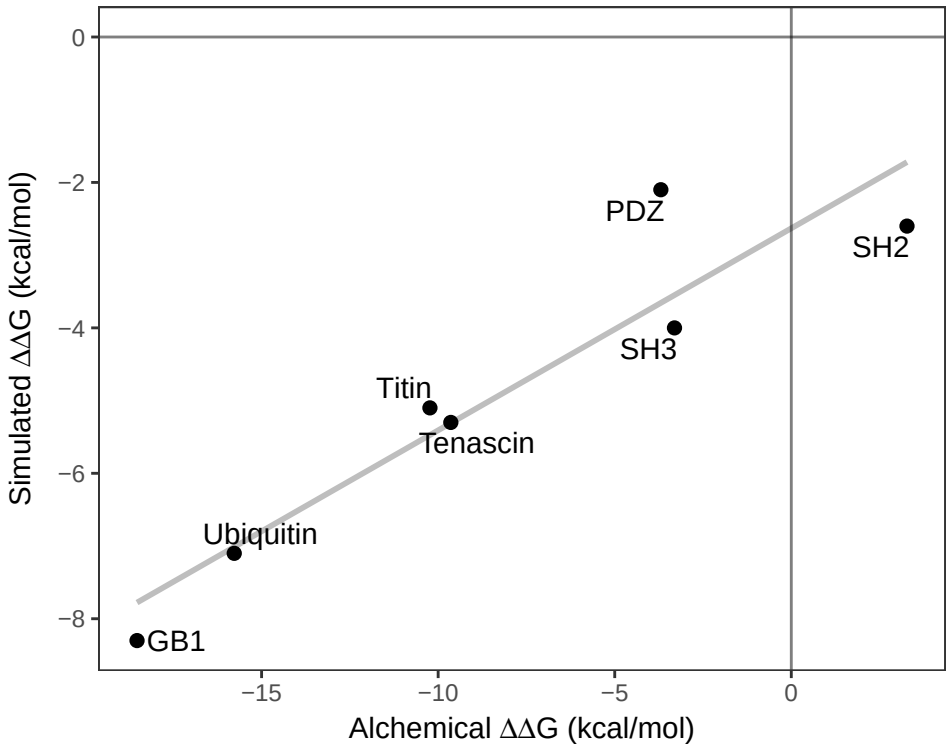

FIGURE 5.4 Correlation between the simulated and alchemical $\Delta\Delta G$. Points correspond to the most stable domain-swapped misfold of each of the seven domains analyzed by Tian and Best (2016): SH3 (PDB:1SHG), SH2 (PDB:1TZE), PDZ (PDB:2VWR), Tenascin (PDB:1TEN), Titin (PDB:1TIT), Ubiquitin (PDB:1UBQ), and GB1 (PDB:1GB1). Figure adapted from supplementary materials in Lafita *et al.* (2018).

Best (2016), although the energy scale differs by a factor of approximately two (Figure 5.4). Domains found as highly similar tandem domain repeats (GB1 and Ubiquitin) are predicted to be more resistant to domain swap misfolding (lower $\Delta\Delta G$), while domains found as tandem domain repeats with low sequence similarity (Titin and Tenascin) and domains found mainly in isolation (SH3, SH2 and PDZ) are predicted to be more susceptible to tandem domain swapping (higher predicted $\Delta\Delta G$).

The direct estimation of the free energy by TADOSS allows the comparison of predicted stability across different domains and is therefore more general than relative probabilities or scores for each domain. Furthermore, the distinction between the two different steps, join and cut, allows rationalising predictions in terms of molecular determinants. For example, as shown in Table 5.1, the linker connecting the domain termini is one of the most significant contributors to the

TABLE 5.1 Alchemical free energy predictions by TADOSS for the representative domains used by Tian and Best (2016). Contributions from each term of the energy function (join and cut) are shown in columns ΔG_J, ΔG_C and ΔG_{Ch} (cut with hinge loop of at least 3 residues). These terms are combined in columns $\Delta\Delta G$ CP, $\Delta\Delta G$ DSD and $\Delta\Delta G$ TDS to estimate the prevalence of each domain folding variant, respectively: circular permutants (CP) involve join and cut, domain swap dimers (DSD) involve hinge loop formation only, and tandem domain swaps (TSD) involve join and hinge loop formation terms. All values are free energies in kcal/mol. Table adapted from Lafita *et al.* (2019).

Domain	PDB	ΔG_J	ΔG_C	ΔG_{Ch}	$\Delta\Delta G$ CP	$\Delta\Delta G$ DSD	$\Delta\Delta G$ TDS
GB1	1PGA	-16.7	3.8	-1.9	-12.9	-1.9	-18.6
UBQ	1UBQ	-13.4	3.8	-2.4	-9.6	-2.4	-15.8
I27	1TIT	-9.7	5.2	-0.5	-4.5	-0.5	-10.2
FN3	1TEN	-7.6	3.8	-2.0	-3.8	-2.0	-9.6
PDZ	2VWR	-5.6	7.8	2.2	2.2	2.2	-3.4
SH3	1SHG	-3.0	5.3	-0.3	2.3	-0.3	-3.3
SH2	1TZE	-1.8	6.7	1.0	4.9	1.0	-0.8

relative stability in GB1 and ubiquitin domains. Hence, for these domains inter-domain linkers are predicted to have a large effect on the formation of tandem domain swaps and could be targeted with protein engineering to reduce the misfolding propensity.

Thanks to its modularity, the alchemical free energy calculation is also applicable to predict circular permutations and domain-swapped oligomers with minor changes in the parameters (Table 5.1). Circular permutations correspond to the operations of joining the termini and forming a hinge loop of length 0 (only cutting the protein chain), while domain-swapped oligomers only require the formation of a hinge loop, since the domains are part of different protein chains and their termini are not connected.

5.2.2 Method validation

In order to validate the TADOSS method, I further compared its predictions with experimental structures and other methods to predict protein circular permutations and domain swap oligomers. Slight variations of the alchemical free energy calculation were used for each case (domain swapping or circular permutation),

as described previously and shown in Table 5.1.

Iwakura *et al.* (2000) characterized the folding units of a DHFR domain by creating all possible circular permutations of a protein and experimentally measuring their stability. The experimental constructs used a linker of five Glycines connecting the domain termini, which would reduce the ΔG_J of joining the termini to a negligible -0.1 kcal/mol according to TADOSS, making the alchemical ΔG_C directly comparable to the experimental ΔG measurements. I used a hinge loop of length 0 for the calculation of alchemical ΔG_C in order to account for the prediction of circular permutations instead of domain swaps.

The alchemical free energy profile is able to recapitulate experimental observations by Iwakura *et al.* (2000) (Figure 5.5). The most stable circular permutation positions (around positions 20, 70, 125 and 145) in the DHFR are all predicted to be stable by TADOSS, despite differences in the scale of the free energy estimations. Several of the experimentally unstable circular permutations also correspond to the minimum alchemical ΔG_C value of 0 kcal/mol.

Predictions by TADOSS also agree with circular permutation predictions by CPred (Lo *et al.*, 2012b), and hinge loop positions predicted by Ding *et al.* (2006). In Figure 5.6, predictions by CPred are compared with normalised TADOSS free energy estimates of viable circular permutations.

In Figure 5.7, hinge loop positions of experimentally determined domain swap dimers and predictions by Ding *et al.* (2006) correspond to maximums of the TADOSS ΔG_C profile. Although the experimentally observed domain swap dimers are not always predicted as the most probable hinge loop position by TADOSS, they are always found in a maximum of the ΔGc profile. Only ΔG_C is used in this analysis to predict domain swap dimers, as no termini joining is required for domain swap oligomers.

5.2.3 Effect of the inter-domain linker

Tian and Best (2016) further observed in their simulations that the length of the inter-domain linker that connects the tandem pair of identical domains has a large effect on the stability of tandem domain swap conformations. Longer and flexible linkers permit the formation of loops that connect the domain termini, disrupting fewer native contacts and making the tandem domain swap more stable by increasing the ΔG_J.

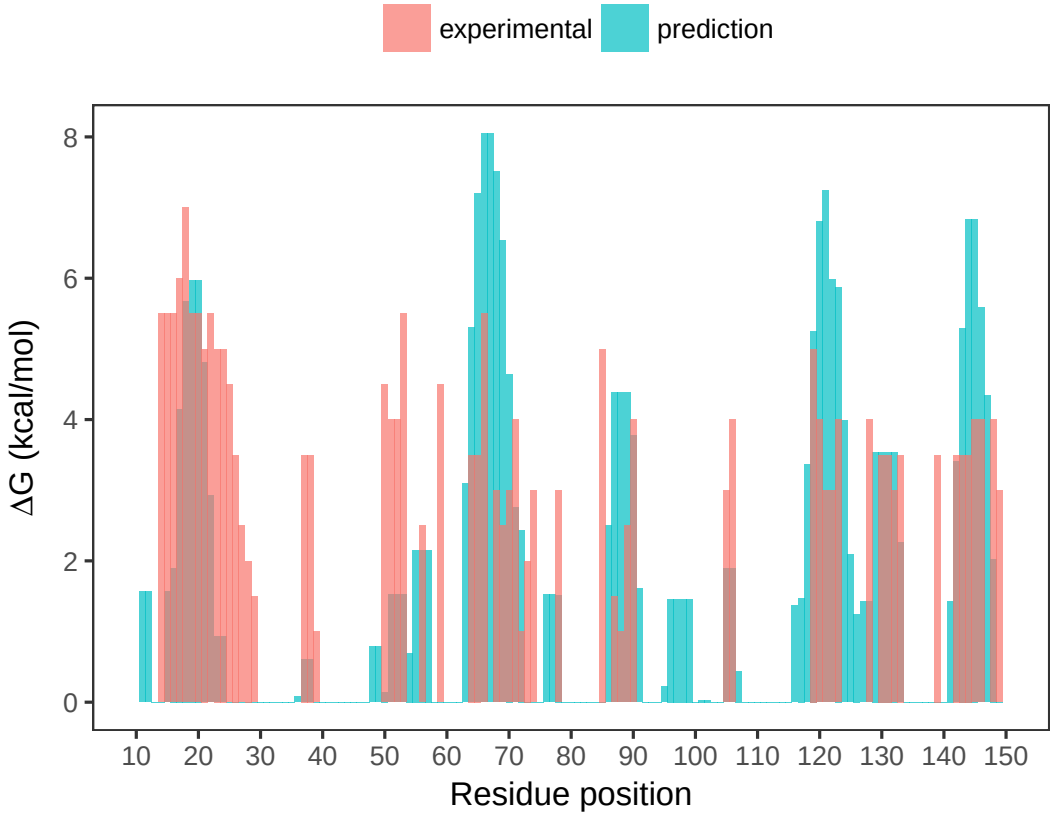

FIGURE 5.5 Comparison of experimental and alchemical ΔG of circular permutations in DHFR. Comparison of the experimental unfolding ΔG of circular permuted constructs of a DHFR domain (PDB:1RX4) measured by Iwakura *et al.* (2000) and the TADOSS alchemical cut free energy (ΔG_C) prediction. Peaks on the ΔG profile indicate structural regions susceptible to viable circular permutations, that is the position that form the new N- and C-termini of the domain. Experimental ΔG is set to 0 in cases where the DHFR circular permutation constructs did not fold into stable structures. Figure adapted from supplementary materials in Lafita *et al.* (2018).

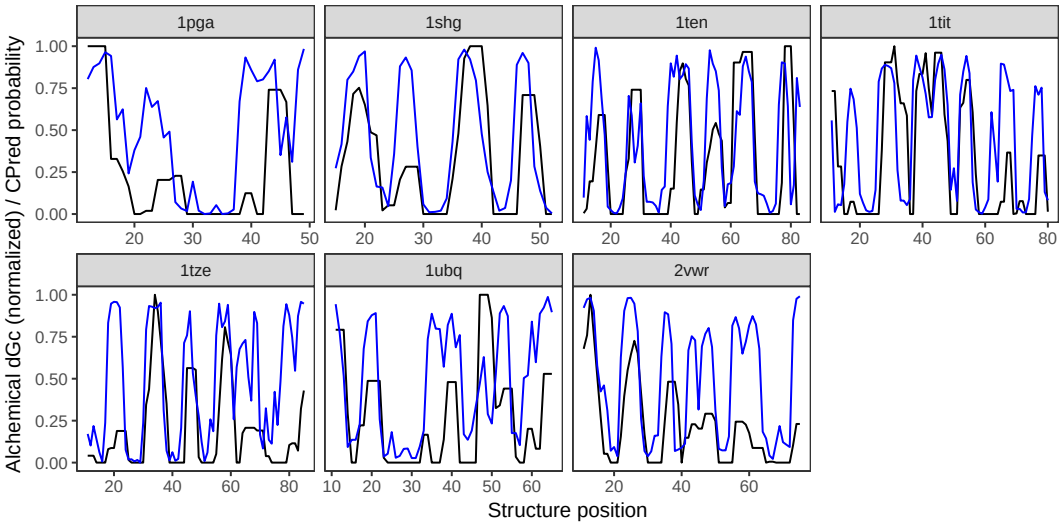

FIGURE 5.6 Prediction of viable circular permutations. Profile predictions by TADOSS (blue) and CPred (black). Alchemical free energies from TADOSS have been normalised in the range 0-1 to be comparable to CPred probabilities. A hinge loop of length 0 was used to account for the prediction of circular permutations. PDB codes for each domain structure are shown as headers for each subplot. Figure adapted from supplementary materials in Lafita *et al.* (2018).

In order to account for the effect of the inter-domain linker length, I introduced a new parameter in TADOSS that reduces the effective distance between the termini of the domain proportional to the length of the inter-domain linker. For example, if the distance between the N- and C-termini of the domain is 14Å, our model would require at least 4 residues to be unfolded from the native structure (each unfolded residue would cover 3.5Å of the distance). By setting the new linker length parameter to one residue, only three other residues would have to be unfolded; and by setting it to four residues, no other residues in the native structure would have to be unfolded, effectively relaxing the constraints to form a tandem domain swap.

I also find good agreement between the effects of the inter-domain linker length predicted by TADOSS and obtained by molecular simulations by Tian and Best (2016) (Figure 5.8). One of the differences I observed is that our alchemical model has an upper limit from which the linker length would always correspond to the same ΔG_J=0 value, while the simulations seem to converge to slightly different values. For example, in Ubiquitin, linker lengths above 10 residues all

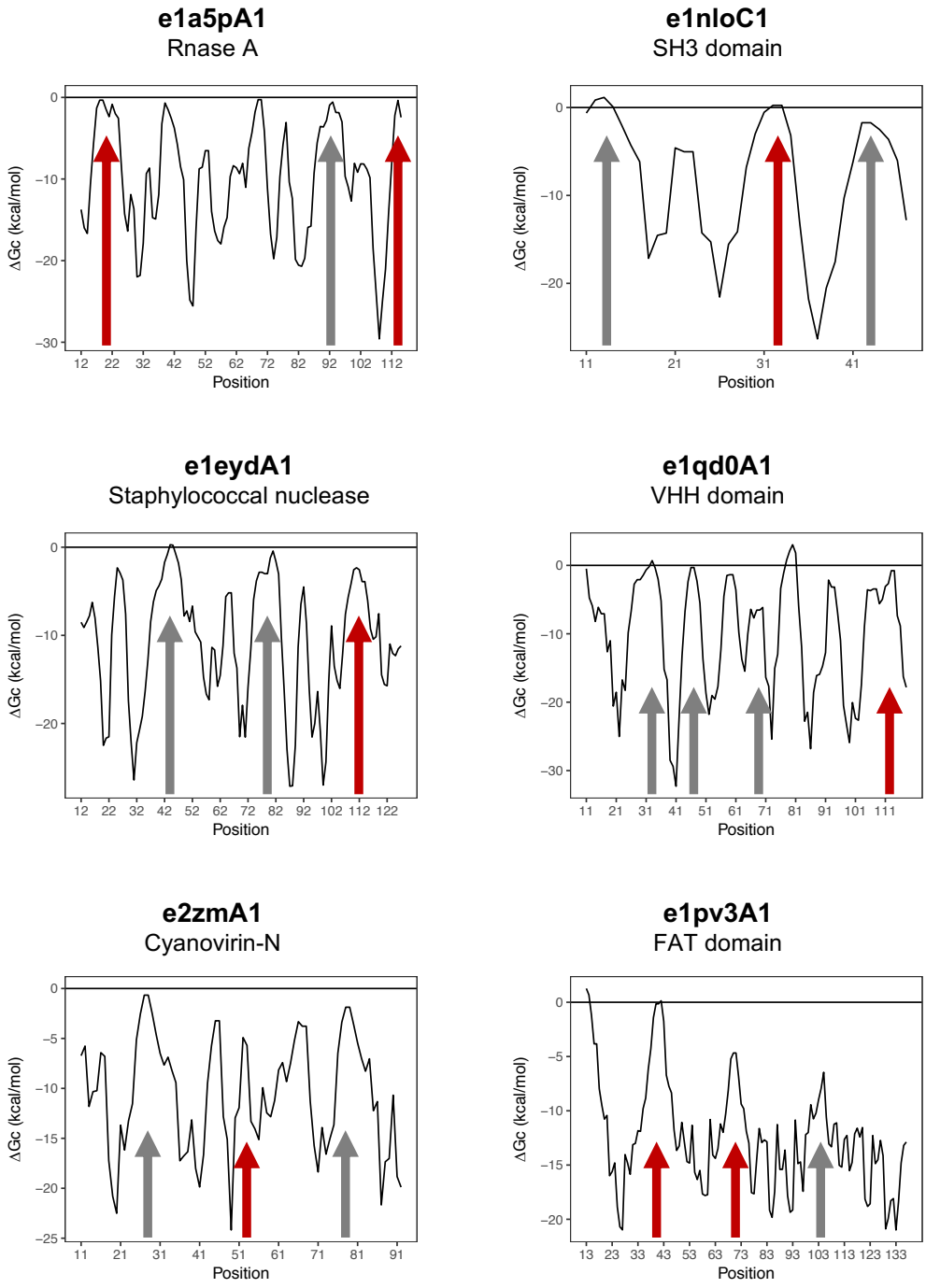

FIGURE 5.7 Prediction of experimental hinge loop regions in the domain examples presented by Ding *et al.* (2006). Experimentally observed domain swap dimers marked as red arrows, and most probable hinge loop predictions by Ding *et al.* (2006) as grey arrows. Domains are annotated with their ECOD identifiers. Figure adapted from supplementary materials in Lafita *et al.* (2018).

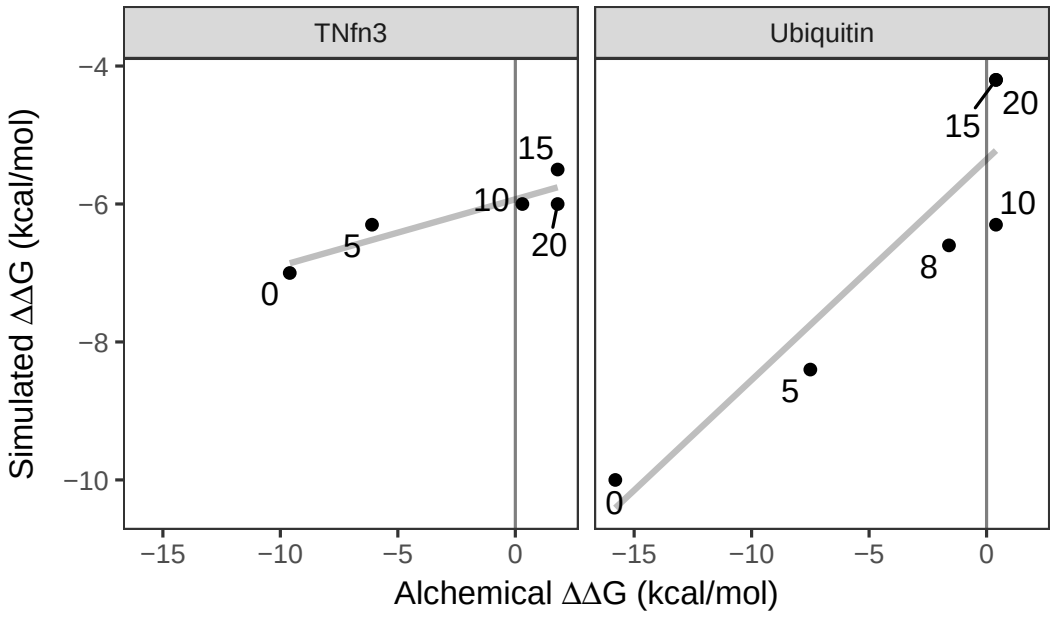

FIGURE 5.8 Inter-domain linker effects on the alchemical $\Delta\Delta G$. Correlation between the $\Delta\Delta G$ from simulations by (Tian and Best, 2016) and from TADOSS predictions for different inter-domain linker lengths (shown as labels) of a Fibronectin type III domain from Tenascin (TNFn3, PDB:1TEN) and a Ubiquitin domain (PDB:1UBQ). Figure adapted from supplementary materials in Lafita *et al.* (2018).

have ΔG_J=0, so longer linkers do not have any additional effect. Intuitively, this can be interpreted as following: once the length of the inter-domain linker is sufficient to cover the distance between the N- and C-termini of the domain, no other residues need to be unfolded and the ΔG_J penalty of forming a tandem domain swap is negligible.

5.2.4 Running time and scalability

Alchemical free energy calculations by TADOSS are very fast. A domain of average length (100 residues) takes less than 2 seconds on a MacBook Pro laptop with a 2.9 GHz processor. The running time scales quadratically with the number of residues in the input domain structure, as shown in Figure 5.9. The calculation for each domain runs on a single separate CPU, so analyses on large scale datasets can also be easily parallelised.

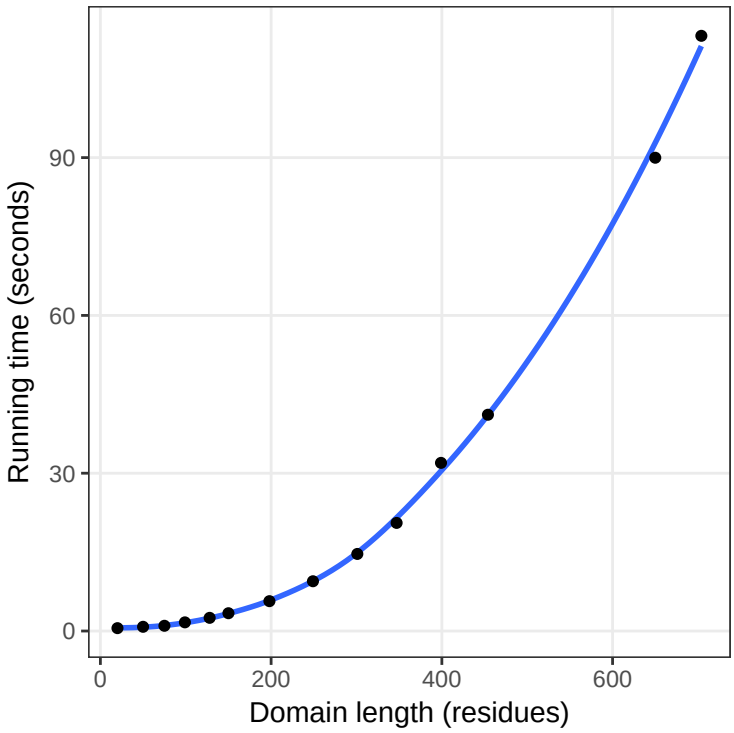

FIGURE 5.9 Running time to calculate the alchemical $\Delta\Delta G$ with TADOSS as a function of the length of the input domain. Calculations performed on random ECOD domains of different lengths, sorted increasingly: e2bl6A2, e1nh2D2, e2egeA1, e3pv5B2, e1c20A1, e2b06A1, e3c6aA1, e4jkxA1, e4g3hA1, e1uqyA1, e4e4jA1, e1ua4A1, e1lkxA1, and e2iukA4. Figure adapted from supplementary materials in Lafita *et al.* (2018).

5.3 Determinants of protein misfolding

I developed TADOSS in order to analyse large datasets of protein domains and find general structural determinants of protein misfolding in tandem domain repeats. Here, I investigate the effect of three structural properties of tandem domain repeats for the misfolding propensity of proteins: domain topology, length and inter-domain linker. I use datasets of domain structures from the ECOD database (Cheng *et al.*, 2014) to compare TADOSS predictions between different secondary structures, topologies and families of domains.

5.3.1 Domain fold and topology

The secondary structure content of a domain does not seem to be important for the misfolding propensity of tandem repeated domains, as both mainly-β and mainly-α domains have been found forming highly similar tandem domain repeats. On the contrary, the domain fold (spatial arrangement of secondary structure elements) and topology (the connectivity of these secondary structures), appear to be important factors. Certain types of topologies, such as Immunoglobulin-like and Ubiquitin-like folds, are commonly found as tandem repeats with high sequence identity, while others are rarely found as repeats or only at low sequence identities.

First, I selected a subset of 129 representative domains[1] from topology groups (T-group) in the ECOD database in order to investigate general patterns among secondary structure classes and domain topologies. I estimated their misfolding propensity as alchemical $\Delta\Delta G$ with TADOSS (Figure 5.10) and I found that a significant proportion of domains (38%) are predicted to be susceptible to tandem domain swapping, with positive peaks of alchemical $\Delta\Delta G$ in their profile, and 18% of the domains have both positive ΔG_C and ΔG_J.

I observe no differences in the distribution of alchemical $\Delta\Delta G$ for secondary structure types, and both α-helical, β-sheet and mixed domains contain topologies predicted to be resistant and susceptible to tandem domain swap formation (Figure 5.10). Among the most resistant topologies (below -10 kcal/mol), I find GB1 and Ubiquitin domains, but also other α-helical bundles, such as "Acyl-CoA

[1] Table of selected domains and alchemical free energy estimates can be found in the GitHub repository: https://github.com/lafita/tadoss

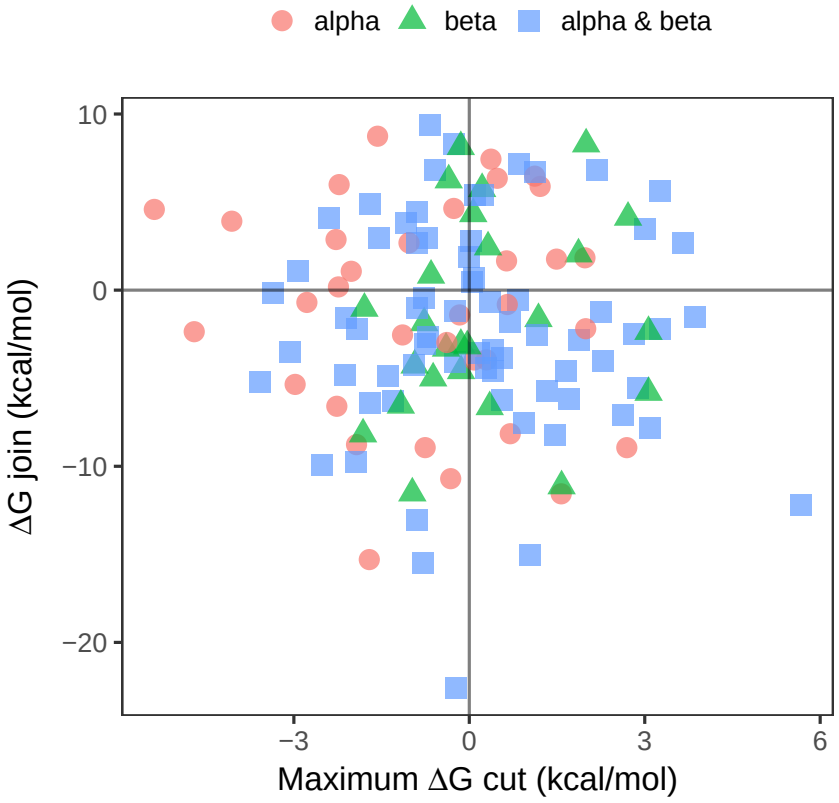

FIGURE 5.10 Alchemical free energy estimations across ECOD topology representative domains. Alchemical ΔG_C and ΔG_J estimated by TADOSS across manual representatives for each T-group (topology) of the ECOD database. Domains are split according to their secondary structure composition: α-helical (alpha), β-sheets (beta), or mixed α-helical and β-sheet (alpha & beta).

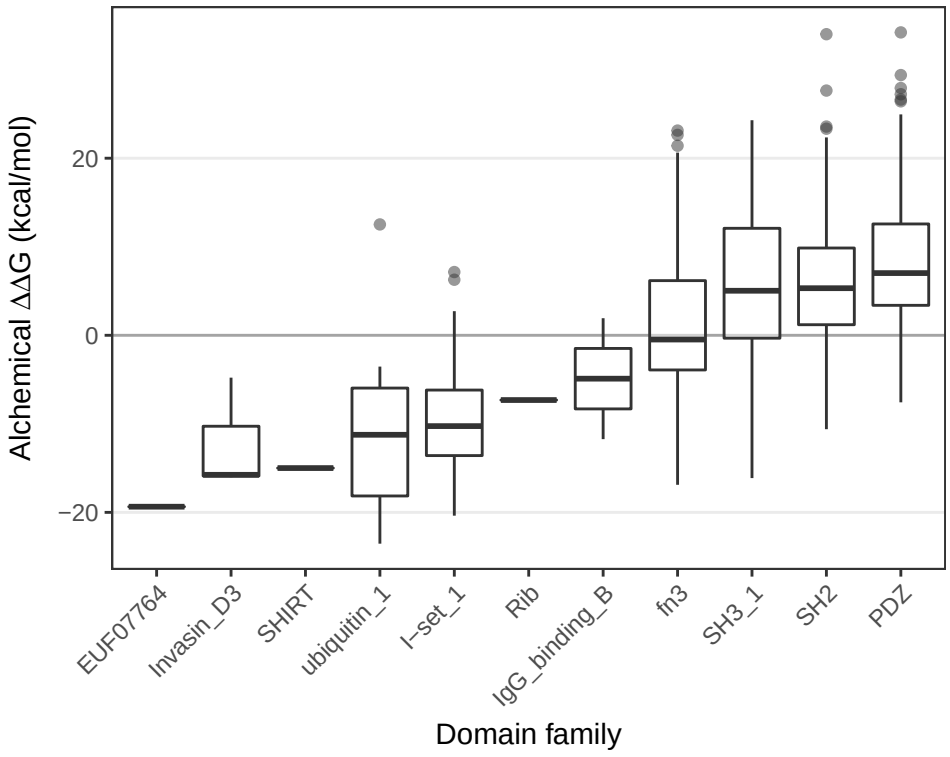

FIGURE 5.11 Tandem domain swap stability predictions for domain structures in different protein families. Box-plot of alchemical $\Delta\Delta G$ estimations for domain structures in ECOD domain families, sorted by their median $\Delta\Delta G$. Higher alchemical $\Delta\Delta G$ values correspond to domains more susceptible to the formation of tandem domain swaps. Domain families commonly found as highly similar tandem repeats: EUF07764 (which corresponds to MBG in Pfam), Invasin_D3, SHIRT, ubiquitin_1, Rib and IgG_binding_B.

binding protein-like" and "YqeY" domains, and α/β three-layered sandwiches. Among the most susceptible topologies (above 4 kcal/mol), I find small β-barrels, such as PDZ and OB-fold domains, other bigger β-barrels and helical topologies such the four-helical up-and-down bundle.

I further calculated the alchemical free energy of ECOD domains within a subset of eleven selected domain families, and compared these predictions to experimental structures of tandem domain repeats: Rib, SHIRT and MBG domains (Figure 5.11). I found a high variability of predictions within domain families, suggesting that small sequence and structural changes to domains can have a large impact on the alchemical $\Delta\Delta G$, but tandem domain repeat structures were predicted to be more resistant to tandem domain swap misfolding.

I observed that one of the most common properties of misfolding resistant domain topologies are the distance and orientation of the N- and C-termini of the domain. Topologies with strongly interacting N- and C-terminal segments, for example forming stable contacts through side chain interactions or β-strand pairings, are more resistant to misfolding; unfolding these highly connected terminal residues would have a high alchemical ΔG_J energy penalty. The N- and C-termini commonly interact through a parallel β-sheet, such as in Ubiquitin and Immunoglobulin folds, resulting in a long distance between the two terminal ends. The formation of the "central domain" of a tandem domain swap conformation would be theoretically less favourable in these domain topologies due to a higher number of unfolded terminal residues required to connect the protein termini.

5.3.2 Domain and loop length reduction

Another misfolding determinant is found at the loops that connect secondary structure elements of protein domains. Shorter loops make more difficult the formation of "hinge loops" connecting the domains of a swapped conformation, while long flexible loops would increase the stability of tandem domain swaps and other misfolded conformations, increasing the ΔG_C. I have previously observed domain length minimisation in tandem domain repeats within protein sequence families, where highly similar adjacent domains were on average shorter than domains of the same family found in isolation.

Here, I studied the effect of the domain length on the alchemical free energy estimates for five widespread domain topologies: PDZ, SH2, SH3, Immunoglobulin-like and Ubiquitin-like domains (Figure 5.12). I found that longer domains from the same topology are predicted to be on average more susceptible to tandem domain swapping than shorter domains in three out of the five topologies; there is only a weak correlation between the alchemical free energy and the domain length for Immunoglobulin-like and SH3 domains.

5.3.3 Inter-domain linker

As shown previously in Figure 5.8, inter-domain linkers are another important structural determinant for the misfolding propensity in tandem domain repeats,

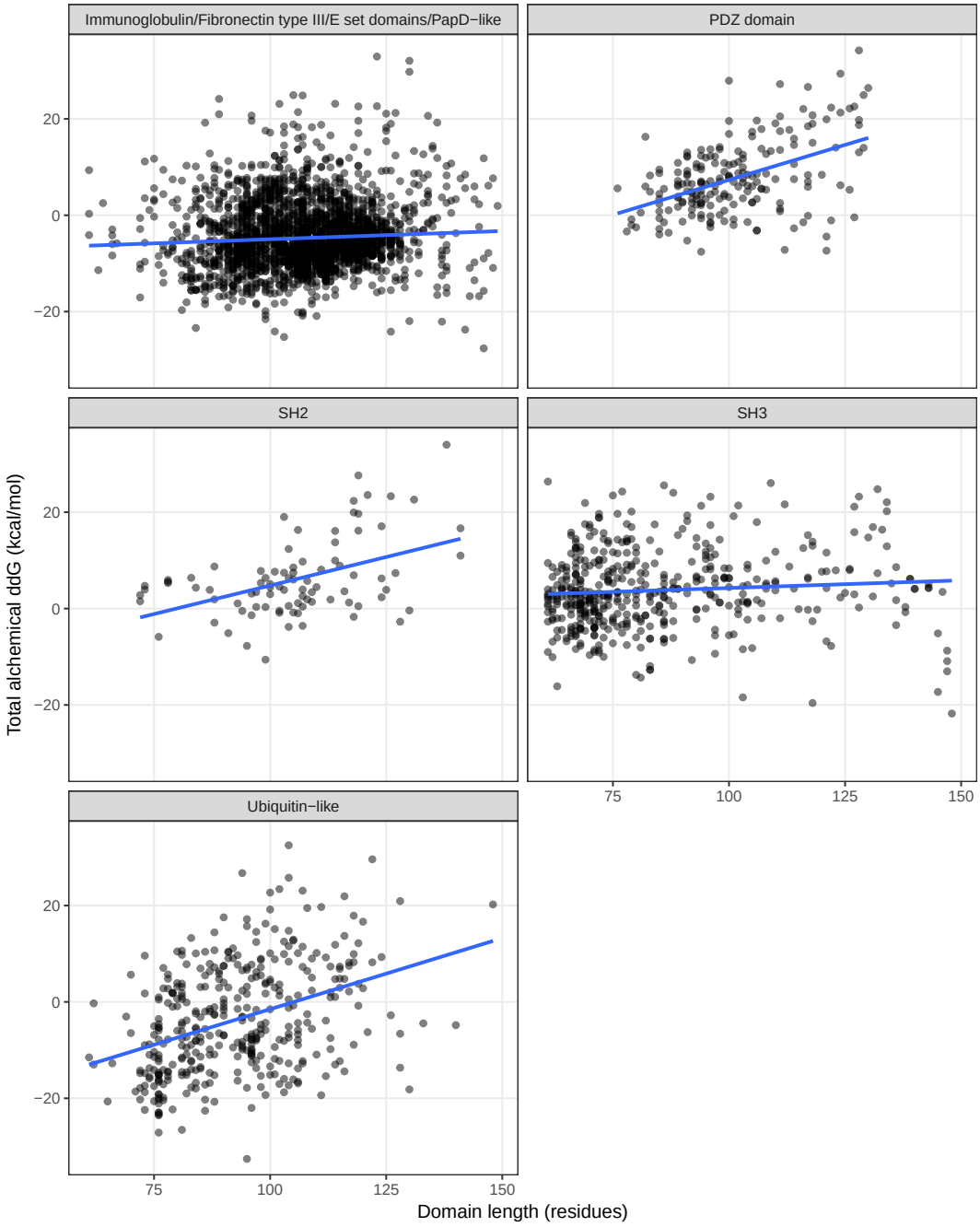

FIGURE 5.12 Effect of domain length on the alchemical free energy. The alchemical free energies of domain structures in five different ECOD topology groups (T-groups) are plotted against their domain length. A linear fit for each point cloud is shown as a blue line.

as they determine the flexibility and orientation of adjacent domains. Short linkers limit the conformational space of the tandem domain pair, while longer more flexible linkers potentially allow alternative conformations with native-like interactions between adjacent domains leading to domain swapping and other intermediate misfolded and transient states.

The length of the inter-domain linker is inversely related to the domain topology in terms of their effect on the alchemical ΔG_J: long distances between domain termini can be compensated by long inter-domain linkers, and vice versa. If the length of the linker is sufficient to cover the distance between the domain termini, no residues would need to be unfolded and the ΔG_J would be negligible in the formation of a tandem domain swap. Structures of pairs of tandem domain repeats have revealed extremely short inter-domain linkers, which contain a high prevalence of Prolines and are predicted to be rigid (Whelan *et al.*, 2019; Whelan *et al.*, 2020), further suggesting that these linkers are relevant structural properties for the misfolding of tandem domain repeats.

5.4 Discussion

In this chapter, I have presented a new method named TADOSS to predict the misfolding propensity of tandem domain repeats via domain swapping. This method is based on the concept of alchemical free energy developed by Tian and Best (2016), and consists in estimating the free energy difference between the native and swap conformations by systematically evaluating the conservation and energy of native contacts in the structure of a domain.

The TADOSS method allowed us to explore the misfolding propensity of diverse sets of protein domain structures, and to identify three domain properties as relevant determinants of tandem domain swap formation: domain topology, domain length and inter-domain linkers. Domain topologies with interacting N- and C-termini, specially those where the termini strands form an antiparallel β-sheet such as β-sandwiches and β-grasp, are more resistant to tandem domain swap formation; but resistant topologies with interacting termini also include helical bundles, for example, and I did not find major differences among secondary structure types.

Shorter domains are also found to be more resistant to misfolding, primarily

due to loop minimisation effects that limit the formation of hinge loops in tandem domain swaps; domain atrophy examples found in tandem domain repeat structures might be an extreme case of this domain length minimisation to resist misfolding and aggregation. Inter-domain linkers further play an important role as misfolding determinants: short and rigid linkers do not allow tandem domain swap conformations, while long linkers do not restrict the conformational space of the tandem domain pair and allow native-like inter-domain interactions.

The work presented in this chapter only addresses one type of misfolding event involving the formation of domain swapping in tandem homologous domains, although other types of misfolding have been observed experimentally. The amount and diversity of available experimental data on the topic is still limited and poses a challenge for theoretical and computational studies such as the one presented here. One of the most evident gaps is the lack of experimentally solved structures of a tandem domain swap. There are sensible reasons for this absence, such as the low prevalence of tandem identical domain repeats in natural proteins, the selective reduction of repeats in crystallization constructs, or the transient nature and heterogeneity of tandem domain swap conformations. However, such a structure would be very helpful to set a precedent for future studies on tandem domain swaps. Nevertheless, the similarity of tandem domain swaps to other better studied protein folding variations, circular permutations and domain swap oligomers, has permitted us to validate the method with experimental structures and compare our predictions to other tools.

On the prediction side, TADOSS only considers the ideal case of two identical domains in tandem, but a more realistic case would be to consider when two domains are highly similar but have some amino acid differences, as the majority of tandem homologous domains found in nature are non-identical. Inclusion of sequence effects will be an important direction for future prediction algorithms, such as simple contact potentials from evolutionary sequence analysis. Future developments could make use of available high-throughput experimental data on the stability of circular permutants, for example. Other features are also important for the formation of tandem domain swaps but have not yet been explored here, such as additional energetically-favorable contacts formed in hinge loops or the rigidity and conformational space of loops and linkers.

Protein misfolding via tandem domain swapping can present major problems

for the design and production of novel protein constructs. For instance, it would be convenient to simply replicate the active domains of a protein to increase avidity of interactions of designed protein biotherapeutics. However, this work suggests these approaches might lead to downstream problems in protein production and activity due to increased protein misfolding and aggregation rates if the sequence is not optimised to avoid tandem domain misfolding. Domain swapping has also been associated with the formation of amyloid fibrils involved in protein deposition diseases in the context of protein oligomerization (Bennett *et al.*, 2006), and similar consequences could derive from tandem domain swap formation.

The misfolding determinants described in this chapter are a step towards a rational description of tandem domain misfolding mechanisms. Improvements in our understanding of this phenomenon might not only improve our understanding of protein folding and misfolding diseases, but also has the potential to improve our ability to produce better and more efficient protein therapeutics.

5.5 Methods

5.5.1 Calculation of alchemical free energy

The structure of the input domain is represented using a coarse-grained structure-based (Gō-like) model, as described by Karanicolas and Brooks (2002). Each residue is approximated to a single point in space, and native interactions between the residues are calculated from their distances and attractive contact energy functions from the Miyazawa-Jernigan matrix (Miyazawa and Jernigan, 1996), resulting in a contact energy matrix of the domain.

From the contact energy matrix, the alchemical free energy profile is calculated by iterating over all possible "cut" positions in the domain. For each position, the energy of unfolding at least three residues to form a hinge loop is calculated from the contact matrix. The domain boundaries are set to the first N-terminal and the last C-terminal residues with at least one native contact to other parts of the domain, effectively ignoring unstructured terminal residues not part of the globular domain. A detailed explanation of the alchemical free energy model and its parameters is included in Appendix C.

5.5.2 TADOSS implementation and availability

TADOSS is written in Python and released as a Bash script with a simple command line interface. The program is fully automated and takes as input the structure of a protein domain in PDB format and generates several output files with the alchemical free energy difference estimates.

The method only requires `BioPython` (Cock *et al.*, 2009), to parse and manipulate the domain structures, and `reduce` (Word *et al.*, 1999), to add Hydrogen atoms. The TADOSS software is open source, documented and available for free under an MIT license on GitHub: `https://github.com/lafita/tadoss`.

Chapter 6

Protein modelling from distance matrices

> *"It is obvious that a partial matrix can be completed to a distance matrix if and only if each connected subset of its associated graph specifies a partial matrix that can be completed to a distance matrix."*
> - Trosset (2000): "Distance matrix completion by numerical optimization"

In this final results chapter, I present a new approach to model protein structures based on Euclidean distance geometry. I use the method to create models of protein structural rearrangements discussed previously in this thesis, including domain atrophy and swapping cases, and I explore further applications of the models to investigate protein geometry and flexibility.

Part of the results presented in this chapter have been published as a Methods article in *F1000 Research*: Lafita and Bateman (2020). I developed the method, performed the analyses and wrote the manuscript.

6.1 Introduction

Structural models are used across a wide range of molecular biology applications, including the study of protein function mechanisms at the molecular level, the interpretation of the effect of variants and mutations in proteins, and the rational design of new proteins. Experimental structure determination techniques provide high resolution atomic models of proteins and other biomolecules, and these models have enabled major breakthroughs in modern molecular biology, such as the discovery of the DNA double helix. Besides their utility as visual representation of macromolecules, structural models are used to derive other energetic and geometric properties of proteins. Experimental models provide a valuable starting point for these types of analyses but they are static snapshots of a protein state; complex analyses require sophisticated modelling techniques to explore the protein conformational space (Brini *et al.*, 2020).

Throughout this thesis I have presented experimental structures of globular domains that form tandem repeats in proteins, such as the Rib domains. Among them, I identified cases of domain atrophy, and I discussed the susceptibility of tandem domains to misfold via domain swapping; both examples can be defined as structural rearrangements. By modelling domain atrophy and swapping events in protein domains, it would be possible to visualise and interpret computational predictions; models can further be used for more sophisticated analyses of protein conformation.

Techniques to generate structural models of proteins are typically based on molecular dynamics simulations that are difficult to scale due to their complicated setup, analysis and computational complexity, and include tools such as Rosetta (Leman *et al.*, 2020) and GROMACS (Abraham *et al.*, 2015). Large structural rearrangements of proteins, such as domain atrophy and swapping, can however be modelled directly using distance matrices, resulting in simpler and more efficient models. Furthermore, this approach to generate 3D models of protein conformations can be used to evaluate other properties of structural rearrangements such as their geometrical feasibility and flexibility.

In the following two introductory subsections, I review cases of structural rearrangements discussed throughout in this chapter and introduce Euclidean distance matrices (EDMs) as a mathematical tool to model the structure of proteins. Next, I describe in detail a protein modelling approach based on EDMs, and

demonstrate various applications to model circular permutations, domain atrophy and domain swapping, revisiting cases discussed previously in this thesis.

6.1.1 Protein structural rearrangements

Most proteins fold into compact globular domains stabilised by hydrogen bonds and a hydrophobic core. These domain folds are conserved throughout evolution, typically allowing for sequence mutations and short insertions and deletions (indels) that only cause local perturbations to the domain core. Certain evolutionary events introduce large structural rearrangements in protein domains that conserve the majority of core interactions. Some examples, shown in Figure 6.1, are domain atrophy, where core secondary structures of a domain are deleted (Prakash and Bateman, 2015); circular permutations, where the sequence order of the domain is permuted, connecting the N and C-termini and introducing new termini at a specific "cut" position (Uliel *et al.*, 2002); and domain swapping, where two independent domains exchange secondary structures to form an intertwined dimer (Rousseau *et al.*, 2012).

These structural rearrangements have been extensively studied both experimentally (Iwakura *et al.*, 2000) and computationally (S. Yang *et al.*, 2004) due to their importance for protein stability and function, and a number of computational methods have been developed over the years to predict domains susceptible to circular permutations and domain swaps (Lo *et al.*, 2012b; Ding *et al.*, 2006; Lafita *et al.*, 2018). However, generating 3D models of these structural conformations proves challenging, often requiring complicated software pipelines and long simulations such as those employed in studies by Ding *et al.* (2006) and Tian and Best (2016).

These large structural changes can be, however, naturally represented as distance matrix transformations, exploiting the conserved native residue contacts at the protein core. Here, I present an alternative modelling approach that consists of representing domains as Euclidean distance matrices (EDMs) and generating rearranged structures by applying a series of matrix operations.

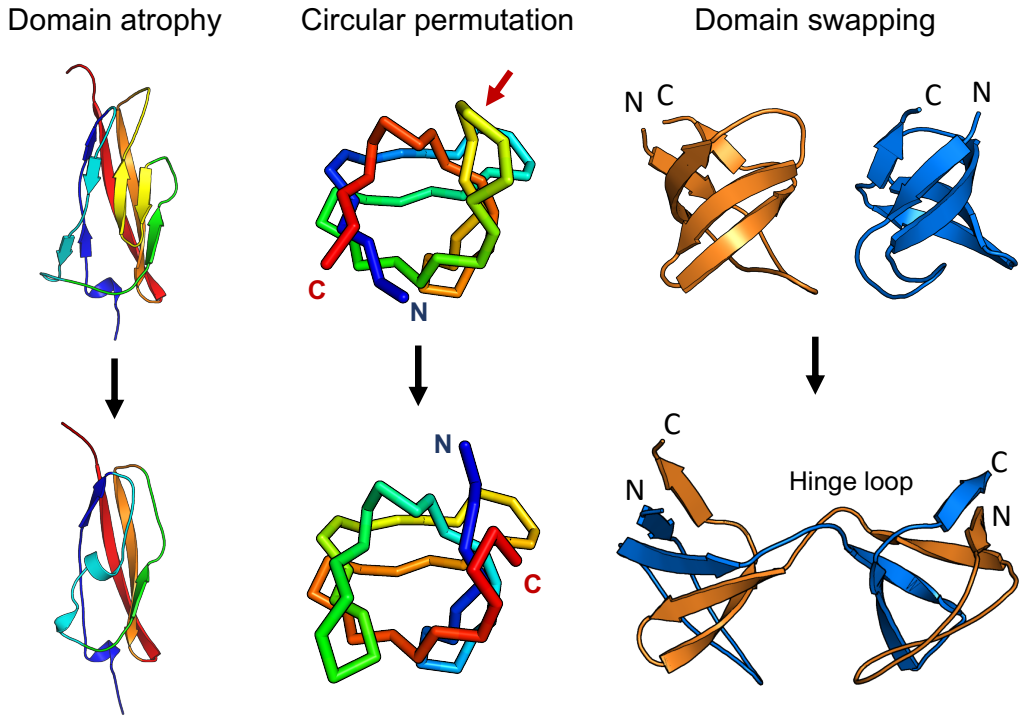

FIGURE 6.1 Examples of structural rearrangements in protein globular domains: domain atrophy in Rib domains (PDB: 6S5W, 6SX1), and a circular permutation and domain swap dimer of an SH3 domain (PDB:1SHG). Structures visualised using PyMol.

6.1.2 Euclidean distance matrices (EDMs)

Euclidean distance matrices (EDMs) are matrices of squared distances between points in an m-dimensional Euclidean space (Dokmanic *et al.*, 2015).

Given a set of n points in three dimensions $p = \{x, y, z\}$, the matrix of points P can be defined as:

$$P = \{p_1, p_2, p_3, \cdots, p_n\} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ y_1 & y_2 & y_3 & \cdots & y_n \\ z_1 & z_2 & z_3 & \cdots & z_n \end{bmatrix} \tag{6.1}$$

and the entries of the Euclidean distance matrix A are the squared distances (Euclidean norm) between points: $a_{ij} = a_{ji} = d_{ij}^2 = \|p_i - p_j\|^2$, arranged as following:

$$A = EDM(P) = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & 0 & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & 0 & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & 0 & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 0 \end{bmatrix} \tag{6.2}$$

EDMs are square and symmetric matrices with zero diagonal and non-negative off-diagonal values, fulfilling the triangle inequality: $\sqrt{a_{ij}} \leq \sqrt{a_{ik}} + \sqrt{a_{kj}}$. Thanks to their many useful properties they have been used for applications in sensor localisation, molecular conformation and dimensionality reduction, among others.

Representing proteins as distance matrices has a number of practical advantages over atomic coordinates. Distance matrices are invariant to point rotations and translations: identical proteins will always have the same distance matrix irrespective of their location and orientation, facilitating their comparison without the need to superpose them in space. However, EDMs are also invariant to mirror images, meaning that a protein and its non-natural mirror image are indistinguishable from their distance matrix.

Algorithms for EDMs have been previously used to model protein structures from distance constraints derived by Nuclear Magnetic Resonance (NMR) experiments (Alipanahi *et al.*, 2013), although MD-based protocols such as CNS (Crystallography and NMR System) (Brünger *et al.*, 1998) are more prevalent

in the field. EDM-based algorithms are particularly well-suited for applications where a large fraction of molecular distances are known with high precision, such as domain swapping and other structural rearrangements where interatomic distances at the native domain core are conserved and can be used as known entries in the matrix.

6.2 Protein modelling using EDMs

Although structural rearrangements such as domain swapping and circular permutation cause large changes in the protein sequence order and the folding topology of domains, core patterns of residue interactions remain conserved and can be represented as distance matrix transformations. Here, I present a protein modelling approach that exploits this observation — that patterns of residue interactions remain unaltered in structural rearrangements — using Euclidean distance geometry operations. This modelling approach aims to be lightweight, flexible and fast compared to simulation-based modelling alternatives, and therefore suitable for large-scale analyses.

A structural rearrangement can be modelled from the structure of a native domain in four steps (6.2A-D). First, the distance matrix of the native domain is computed from its structure; next, the entries of the matrix are rearranged to represent the structural changes; then, missing entries in the matrix are completed using standard algorithms for Euclidean distance matrix completion; and finally, the distance matrix is inverted back to a 3D model of atomic coordinates. In the following subsections, I describe in more detail each of the different steps of the EDM modelling approach.

6.2.1 Calculation of protein distance matrices

Given an input protein structure, its distance matrix can be computed as the Euclidean norm between all pairs of atoms (Equation 6.2). The distance matrix of a protein with n atoms will therefore be n columns by n rows for a total of n^2 entries.

The granularity of the distance matrix can be easily adjusted by changing the number and type of atoms included from each protein residue. For example, a coarse-grained representation of a protein can be achieved using only C_α atoms.

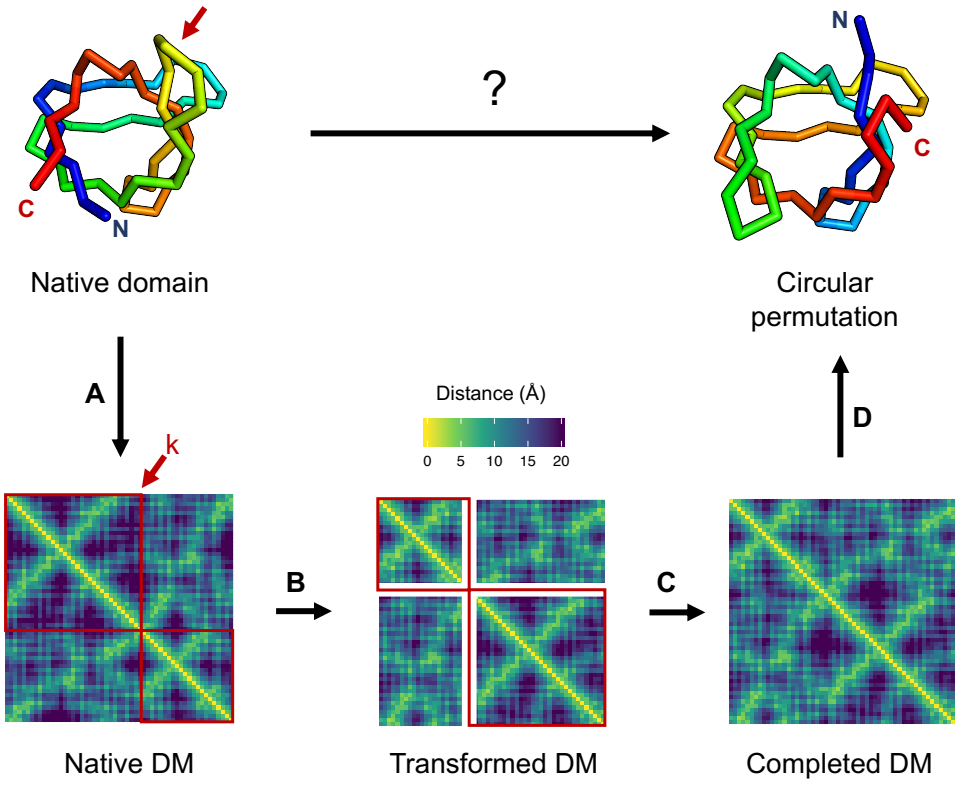

FIGURE 6.2 Protein modelling approach based on Euclidean distance matrices (EDMs). A) Atomic coordinates of the native SH3 domain structure (PDB:1SHG), top left, are converted into a distance matrix (DM), shown at the bottom left as a heatmap of C_β distances in a viridis color scale. B) A matrix transformation operation corresponding to the structural rearrangement is applied to the distance matrix (bottom middle), here shown as a red arrow for the circular permutation cut position k and red boxes encapsulating permuted regions of the matrix. C) Values for unknown entries in the transformed distance matrix, shown as white stripes, are filled using an EDM completion algorithm to generate a complete EDM (bottom right). D) The distance matrix is finally converted to points in 3D corresponding to the atomic coordinates of the rearranged domain structure (top right). Figure adapted from Lafita and Bateman (2020).

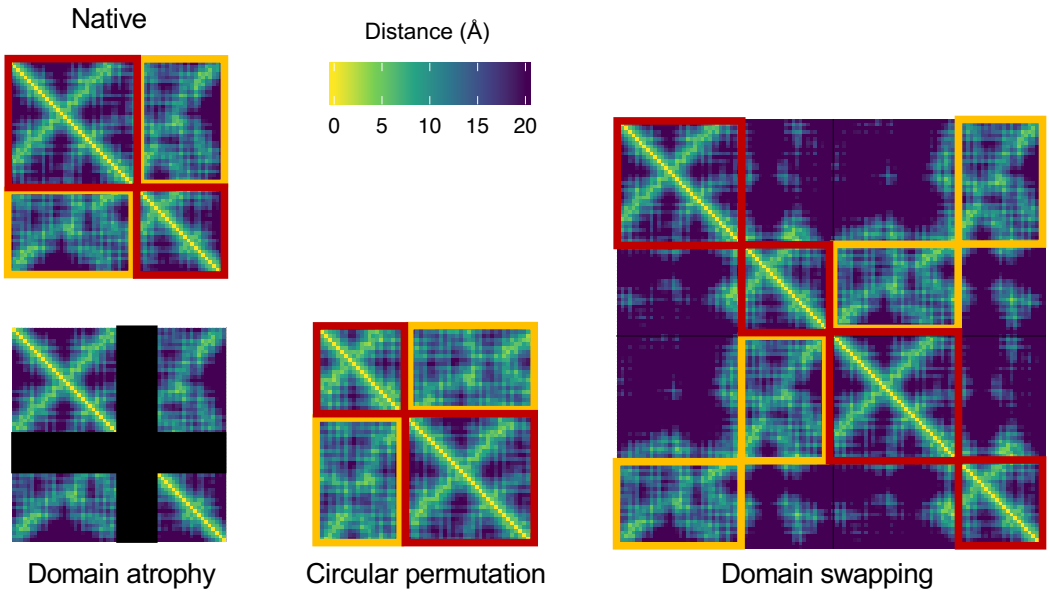

FIGURE 6.3 Schematic representations of distance matrix transformations for three structural rearrangements: domain atrophy, circular permutation and domain swapping. Distance matrices (C_β atoms) of native EPS8 SH3 domain (PDB: 1I0C) and models for domain atrophy, circular permutation and domain swaps generated with EDMs. Conserved intra-domain contacts in the native distance matrix are highlighted as red boxes, while contacts moved to inter-domain (domain swap) and permuted (circular permutation) are highlighted as orange boxes. Rows and columns of the distance matrix deleted in domain atrophy are highlighted as black rectangles.

Given the large structural changes discussed here, models are typically coarse-grained using C_α or backbone atoms only.

6.2.2 Distance matrix transformations

Protein distance matrices can be transformed by rearranging entries and inserting and deleting rows and columns to represent different structural changes. Here I describe briefly how the distance matrix of a protein can be altered for three cases: domain swapping, circular permutation and domain atrophy. More details about the index rearrangements can be found in the Methods section.

The simplest case is domain atrophy. A structural deletion in a domain can be represented as deleting all columns and rows of the corresponding region in the native distance matrix (Figure 6.3).

In a circular permutation the core of the domain remains unchanged, but the residue connectivity changes. Atomic distances at the N- and C-terminal of a "cut" position are therefore interchanged, as shown in Figure 6.3.

For domain swapping, long-range interactions between residues at each side of the "hinge loop" are moved to be inter-domain, while the other local residue interactions are kept intra-domain (Figure 6.3). The distance matrix of a single native domain can be used to estimate contact matrix regions that are preserved, those that are shifted from intra-domain to inter-domain, and those that would be disrupted upon formation of the domain swap, enabling the construction of the domain swap distance matrix.

6.2.3 Completion of protein EDMs

Partial or incomplete EDMs, where only a subset of distances in the matrix are known, are a common problem in many applications, including molecular modelling. The EDM completion (EDMC) problem is an optimisation problem that consists in finding values for missing entries, while preserving the EDM properties. EDM completion is a convex optimisation problem for an embedding space of unrestricted dimensions, but it is non-convex and NP-hard for restricted dimensions such as molecular conformations in 3D (Drusvyatskiy *et al.*, 2017). However, methods for finding local solutions in restricted dimensions often find global solutions and are practically tractable (Trosset, 2000).

Several EDMC algorithms have been developed, formulated as optimisation problems with different loss functions and constraints, including semidefinite programming (SDP) (Alipanahi *et al.*, 2013) and numerical optimization approaches (Trosset, 2000). The dissimilarity parameterization formulation (DPF) algorithm (Trosset, 2000) is a suitable alternative for protein modelling: the dimensions of the embedding space and upper and lower bounds for unknown distances are built into the loss function, and can be set as input parameters to generate atomic coordinates in three dimensions and use empiric protein atomic distance bounds to avoid atom clashes and unrealistic conformations in the models. The DPF algorithm was also chosen in this study because there is an openly available implementation of the algorithm that can be easily used through a programming interface.

EDMC algorithms converge to an exact solution with zero loss if the resulting

matrix is an EDM; if converging to an exact solution is not possible, algorithms return a solution with minimal loss corresponding to the matrix closest to an EDM. In some cases, there are multiple exact solutions to the completion problem and algorithms will randomly converge to one of the valid solutions. For the protein modelling task, EDMC convergence indicates that a geometrically valid conformation, or multiple valid conformations, exist for the input distance matrix; non-convergence indicates that no conformation would be geometrically possible.

In order to obtain realistic protein conformations in the models, I extracted atomic distances from experimental structures and constructed a lookup table of upper and lower bounds for unknown entries in protein distance matrices. More details are given in the Methods section at the end of the chapter. Given a partial protein distance matrix, lower and upper bounds for unknown distances are set to the corresponding values in the lookup table.

The average, upper and lower bounds for C_α and C_β atom distances in the lookup table are shown in Figure 6.4A. As expected, the uncertainty around the average distance increases with the sequence separation. The lower bound, however, reaches a minimum value of 3.70Å for C_α and 3.53Å for C_β contacts. Values below 3Å are observed for C_α atoms of adjacent residues (separation equal to one), which correspond to cis-Proline conformations. In order to simplify the lookup table of atomic distance bounds, I disregarded cis-Prolines. A subset of the final distance bounds lookup table for two adjacent residues is shown in Figure 6.4B. Dark colours in the table represent distances with low variation, for example those between atoms that form the peptide bond plane (CA0, C0, O0, N1, CA1).

6.2.4 Reconstruction of atomic coordinates

Distance matrices can be inverted into a set of points in an embedding space of m dimensions, three in the case of proteins, using a standard technique known as multidimensional scaling (MDS). Distance matrix inversions can be challenging for noisy and incomplete matrices, but due to the degeneracy in low-rank protein distance matrices, where the number of points is much higher than the number of dimensions, MDS can handle small levels of noise and still produce highly accurate reconstructions. If the MDS point reconstruction results in a non-natural mirror image of a protein, its natural stereoisomer can be obtained by simply changing

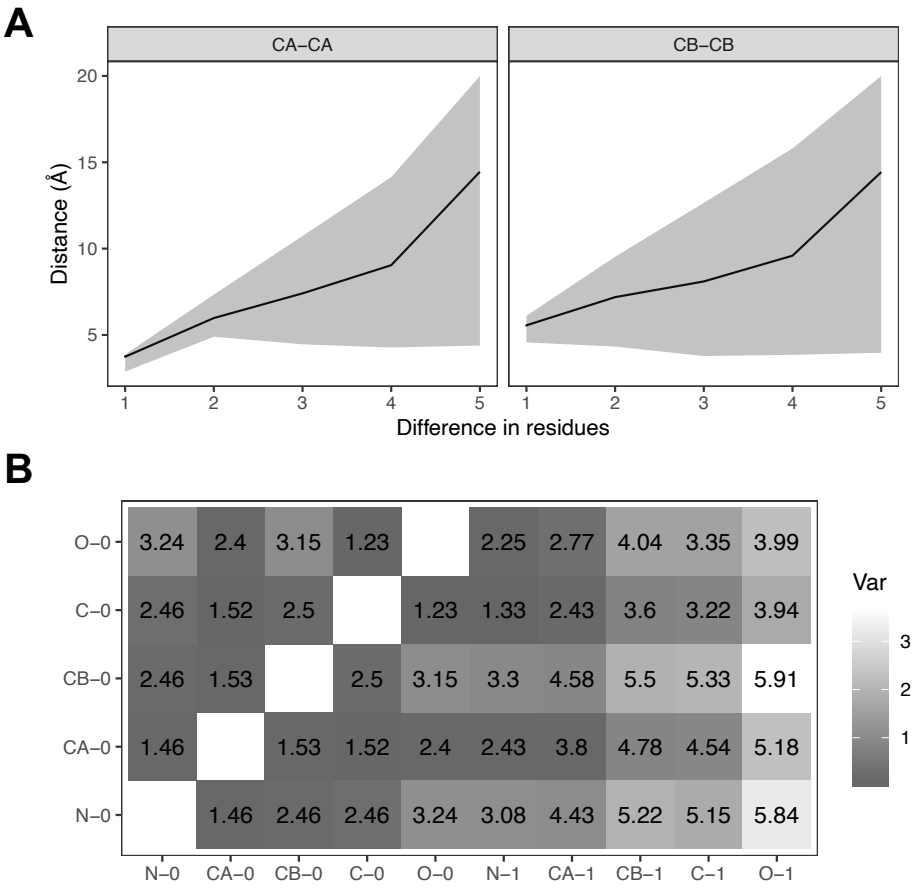

FIGURE 6.4 Atomic distance bounds extracted from experimental domain structures. A) C_α (CA) and C_β (CB) interatomic distances as a function of residue separation along the peptide chain. The average distances are shown as a black line with a grey area for lower and upper bounds. Upper bounds trimmed at 20Å . B) Table of average backbone interatomic distances between adjacent residues (0,1), excluding cis-Prolines. Grey tones indicate amount of variability as the difference between upper and lower bounds in Å, with darker tones for smaller variations. Figure adapted from Lafita and Bateman (2020).

the sign of the z dimension of all points.

6.2.5 Running time

The size of a distance matrix scales quadratically with the number of atoms in the input structure, meaning that the running time and memory of EDM algorithms is expected to be at least $\mathcal{O}(n^2)$. Calculation of distance matrices and multidimensional scaling operations are very fast, and take less than a second for matrix sizes of typical protein domains. The matrix completion step is, however, the bottleneck, varying from just a few seconds to a few minutes, and depends heavily on the size of the input matrix, the number and complexity of unknown entries, and a random convergence factor.

Running times for modelling structural rearrangements of several domain structures are shown in Figure 6.5. C_α models take less than a minute for up to 200 residues, which is within the size of an average protein domain and suitable for large-scale analyses. C_α models of domains up to 400 residues run in less than ten minutes. In comparison, backbone models do not scale particularly well, with an expected factor of 25 times slower than C_α models due to the additional five atoms per residue; backbone models for domains over 200 residues are possible but not practical for most applications.

These running times were calculated on a single CPU with 8GB of RAM. The small computational resources needed to create a model allows one to run multiple modelling jobs concurrently for large-scale analyses.

6.2.6 Experimental validation

To validate EDM models of structural rearrangements, I performed a qualitative comparison against experimentally determined structures.

The experimental structure of a circular permutation of the YibK methyltransferase protein contains five extra residues at the termini that form a linker loop (Chuang *et al.*, 2019). The distance between the N and C-terminal residues in the native domain is over 20Å, so joining the domain termini would not be geometrically possible without a linker, and the EDM model for the circular permutation does not converge to a solution. Modelling the same circular permutation using two to four unfolded terminal residues do not converge to a valid solu-

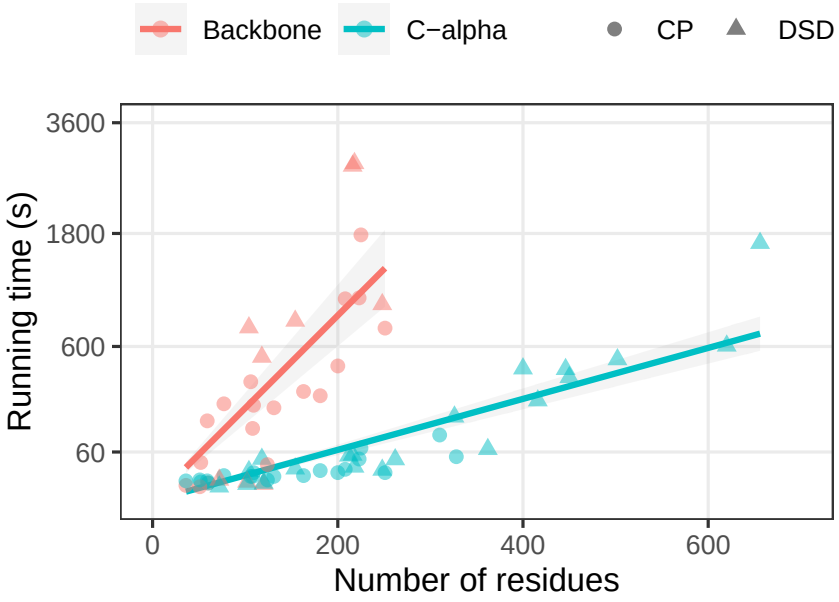

FIGURE 6.5 Running time scaling of protein modelling using Euclidean distance matrices as a function of domain size. Time calculated for backbone and C_α models of circular permutations (CP) and domain swap dimers (DSD) of the 20 ECOD representative domains. Backbone models for domains over 300 residues are omitted. Quadratic fits of the form $y = Ax^2$ for each model granularity are shown as lines with uncertainty grey shades. Running time shown as a square root scale.

tion either, with large loss values and distance errors in the terminal loop region above 8Å and 2Å respectively. Using six unfolded terminal residues converges to an exact solution (Figure 6.6A).

The structures of Cyanovirin-N and EPS8 SH3 domains have been experimentally determined both as monomeric and domain swapped dimers. The EDM model of the Cyanovirin-N domain swap structure from its monomer closely resembles the experimental domain swap (F. Yang *et al.*, 1999) (Figure 6.6B). However, the EDM model of the EPS8 domain swap structure from its monomeric form converges to a valid solution but has a different domain orientation to the experimental domain swap (Kishan *et al.*, 2001) due to its higher inter-domain flexibility (Figure 6.6C). The relative orientation between Cyanovirin-N domains is constrained and cannot change without disrupting natural protein geometry, while EPS8 domains have a greater range of possible orientations.

The domain atrophy in Rib domains involved the loss of a pair of β-strands at the core of its β-sandwich topology (Whelan *et al.*, 2019). The distance between the two ends of the deleted segment is, however, close enough so that a short linker can connect them, as captured by the EDM atrophy model shown in Figure 6.6D. However, the atrophy event causes part of the domain core to be exposed to the surface. In the Rib domain, this gap is filled by another part of the protein which adopts a helical turn conformation, a structural compensation common in domain atrophy and that cannot be predicted by the EDM model. However, the EDM model provide a useful starting point to investigate structural consequences of domain atrophy, as the exposed domain core would evidently be a stability problem for the domain.

6.3 Applications of protein EDM modelling

Besides the generation of 3D models for visual inspection and analysis, the EDM protein modelling approach can be further used to systematically investigate protein geometry and flexibility. Here I present two applications for the study of protein misfolded conformations (domain swaps) and structural deletions (domain atrophy).

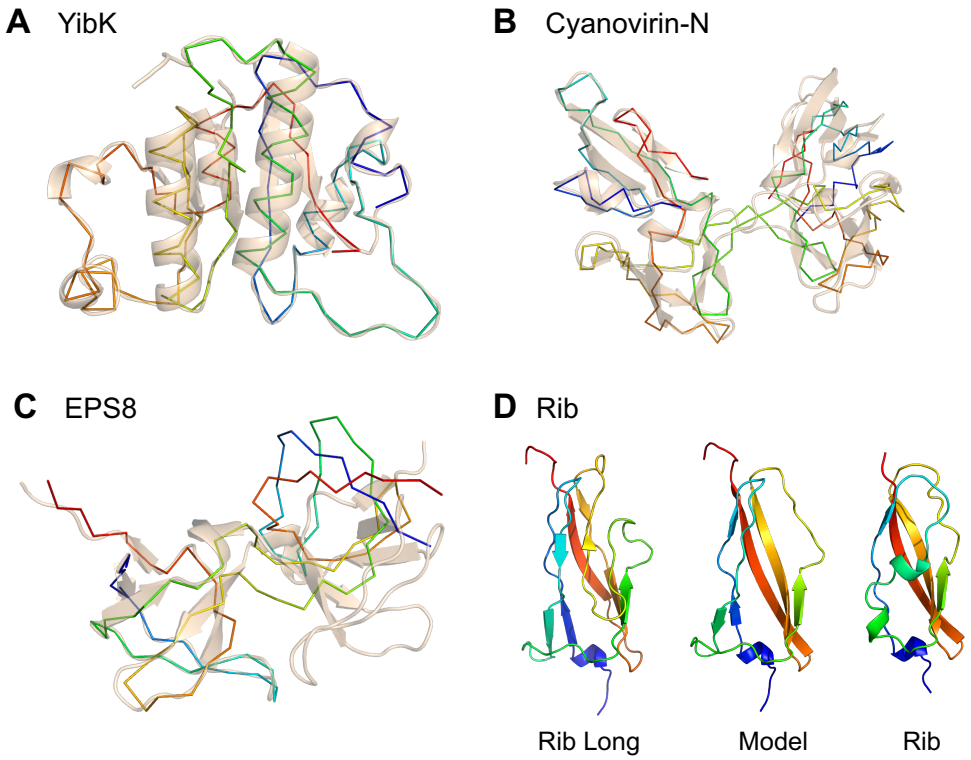

FIGURE 6.6 Models of structural rearrangements generated using Euclidean distance matrices (EDMs). A) C_α model of the circular permutation of YibK methiltransferase as a rainbow backbone, superposed to the native structure (PDB: 1MXI) as transparent cartoon; B) Domain swap dimer model of Cyanovirin-N (PDB: 2EZM) as a rainbow backbone, superposed to its experimental domain swap dimer (PDB: 3EZM) as transparent cartoon; C) Domain swap dimer model of EPS8 SH3 domain (PDB: 1I0C) as a rainbow backbone, superposed to the left-most domain of its experimental domain swap dimer (PDB: 1I07) as transparent cartoon; D) Backbone EDM model of domain atrophied Rib domain (PDB: 6SX1) from the structure of Rib Long (PDB: 6S5W). Structures visualised using PyMol.

6.3.1 Valid domain swap conformations

The TAndem DOmain Swap Stability predictor (TADOSS) method described in the previous chapter reports estimates for the most stable "hinge loop" positions and lengths in a domain, so EDM modelling can be directly used to generate 3D models. This feature has been implemented in the latest version of the TADOSS method, and it not only improves the visualisation and interpretability of predictions, but offers a new avenue to improve its accuracy by discriminating geometrically valid from invalid domain swap conformations.

I systematically modelled all possible domain swap conformations for a subset of six representative domains, and identified geometrically valid and invalid domain swap positions (Figure 6.7). I compared the models with TADOSS alchemical free energy predictions (based solely on protein energetics) to identify domain swap conformations that are both energetically favorable and geometrically possible.

Results show that, despite some positions in the domain are predicted by TADOSS to be susceptible to misfolding via domain swapping, not all of the energetically stable conformations are geometrically feasible and therefore likely to occur experimentally. For example, in the titin I27 domain (Figure 6.7D), there are three energetically favorable domain swaps (positions 30, 40 and 55) but only one of them (position 40) can be modelled with EDMs. In the case of the G5 domain (Figure 6.7C), most of the domain swap conformations cannot be modelled, including the two most energetically favorable at positions 26 and 51, suggesting that domain swapping in G5 domains is geometrically challenging.

6.3.2 Systematic structural deletions

I further used EDM modelling to understand domain atrophy events in proteins by systematically modelling, using EDMs, all possible structural deletions in the ancestral Rib (Rib Long) and G5 domain structures, in order to identify geometrically valid structural deletions (Figure 6.8 and 6.9).

I observed that, for both the Rib and G5 domains, valid structural deletions correspond to regions around β-hairpin turns, which in Figures 6.8 and 6.9 form bands perpendicular to the diagonal very similar to anti-paralel β-strand patterns in distance matrices. The natural domain atrophy observed in Rib and SasG_E

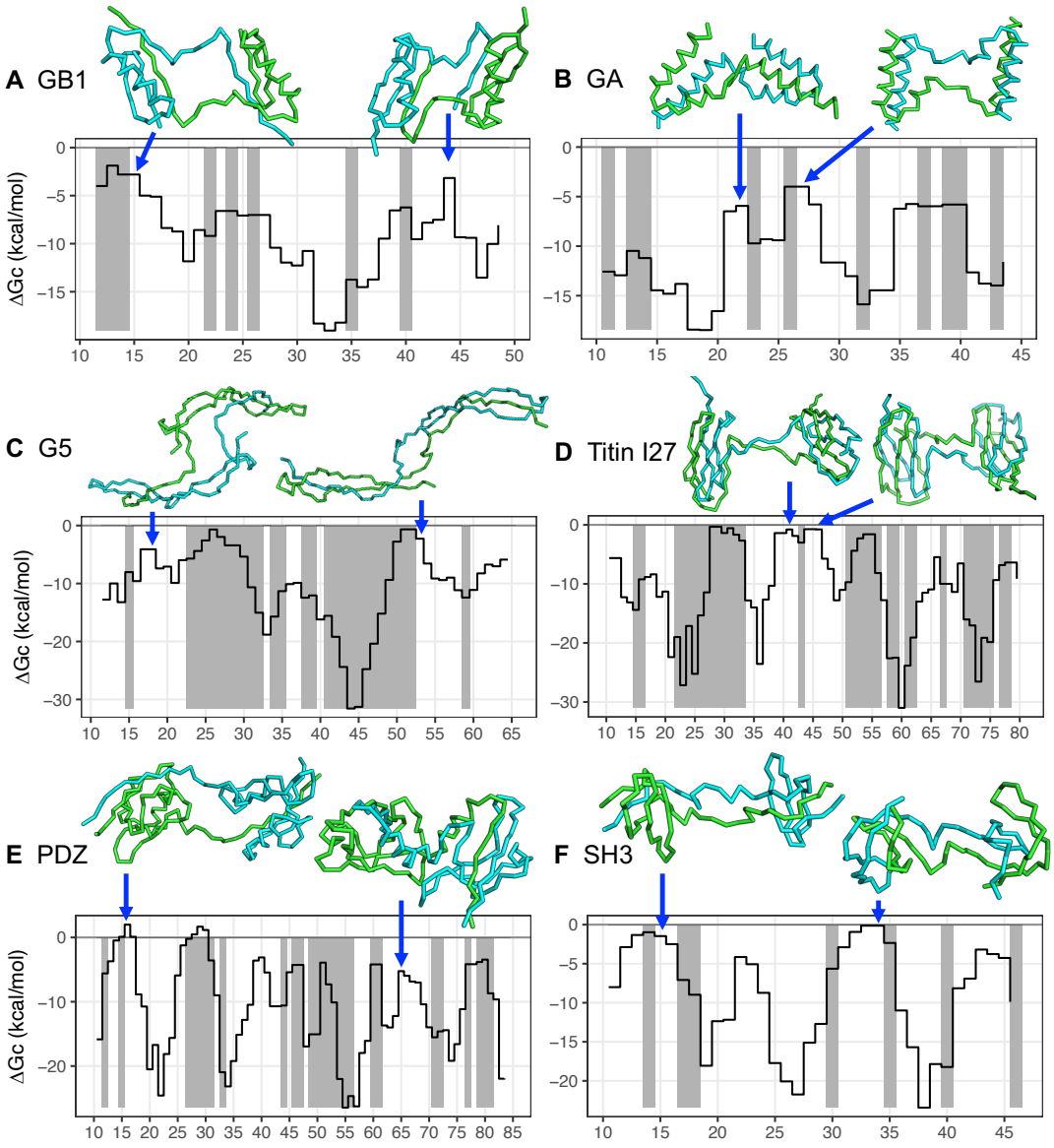

FIGURE 6.7 Discrimination of geometrically valid domain swap conformations by EDM modelling. Invalid domain swap positions (EDM modelling did not converge) are highlighted as grey bars behind the alchemical free energy profile predicted by TADOSS for six globular domains: GB1 (PDB:1PGA), GA (PDB:2J5Y), G5 (PDB:3TIQ), titin I27 (PDB:1TIT), PDZ (PDB:2VWR), and SH3 (PDB:1SHG). Domain swap dimer models for two of the most energetically stable and geometrically valid positions of each domain are shown at the top of each plot, one domain colored in cyan and the other in green.

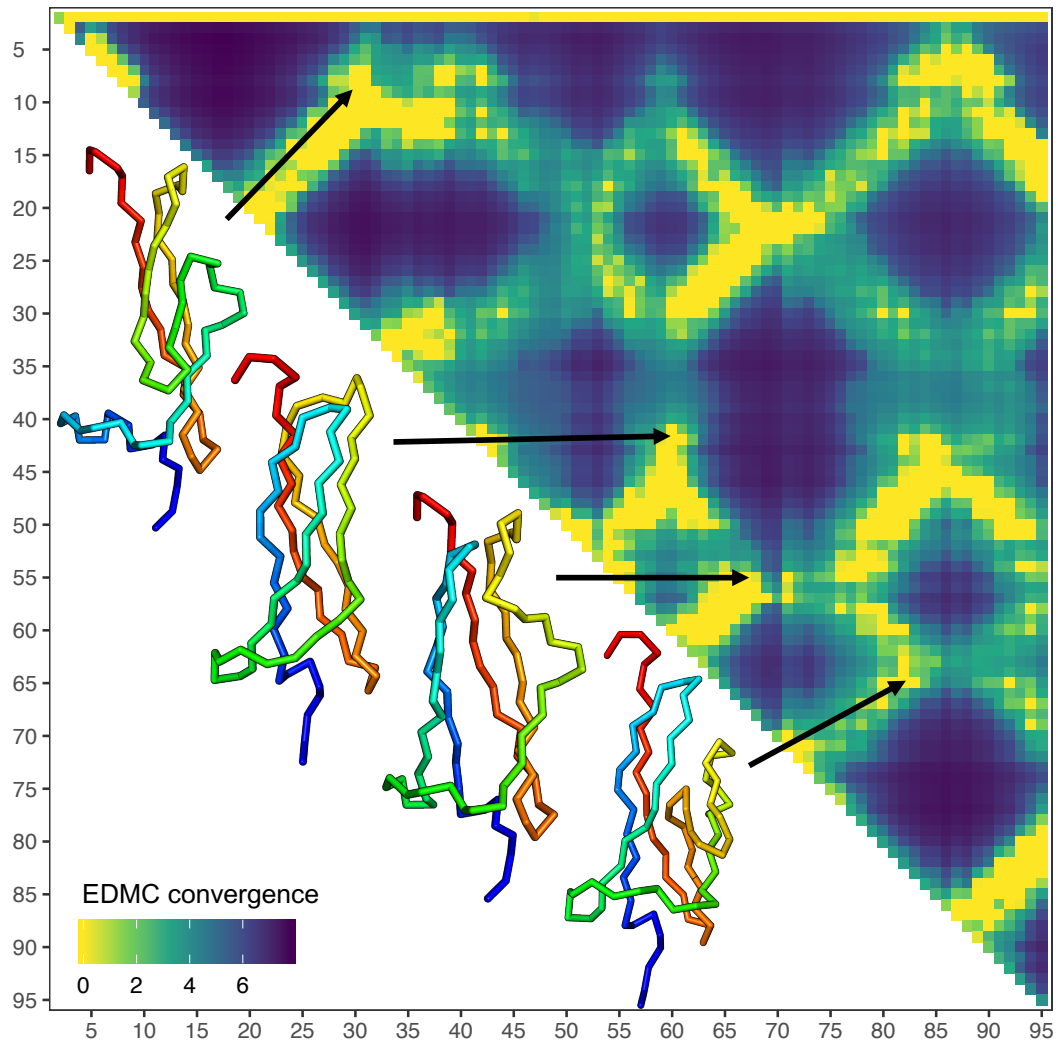

FIGURE 6.8 Modelling systematic structural deletions in the Rib long domain using EDMs. All possible structural deletions in the Rib Long structure (PDB:6S5W) were modelled with EDMs. Convergence values are shown in the upper diagonal part of the matrix: each entry corresponds to a deleted region in the domain from start (row number) to end (column number). Dark blue regions correspond to deletions with high model errors, suggesting they are geometrically invalid, and light yellow entries correspond to error-free models. Four valid models are shown and mapped to their corresponding deleted region: 11-31 (deletion of strands A and B), 41-60 (strands C and D), 55-66 (strands D and E), and 67-80 (strands E and F). The domain atrophy region observed in the Rib domain structure (PDB:6SX1) corresponds to the 55-66 deletion, involving the β-hairpin turn of strands D-E.

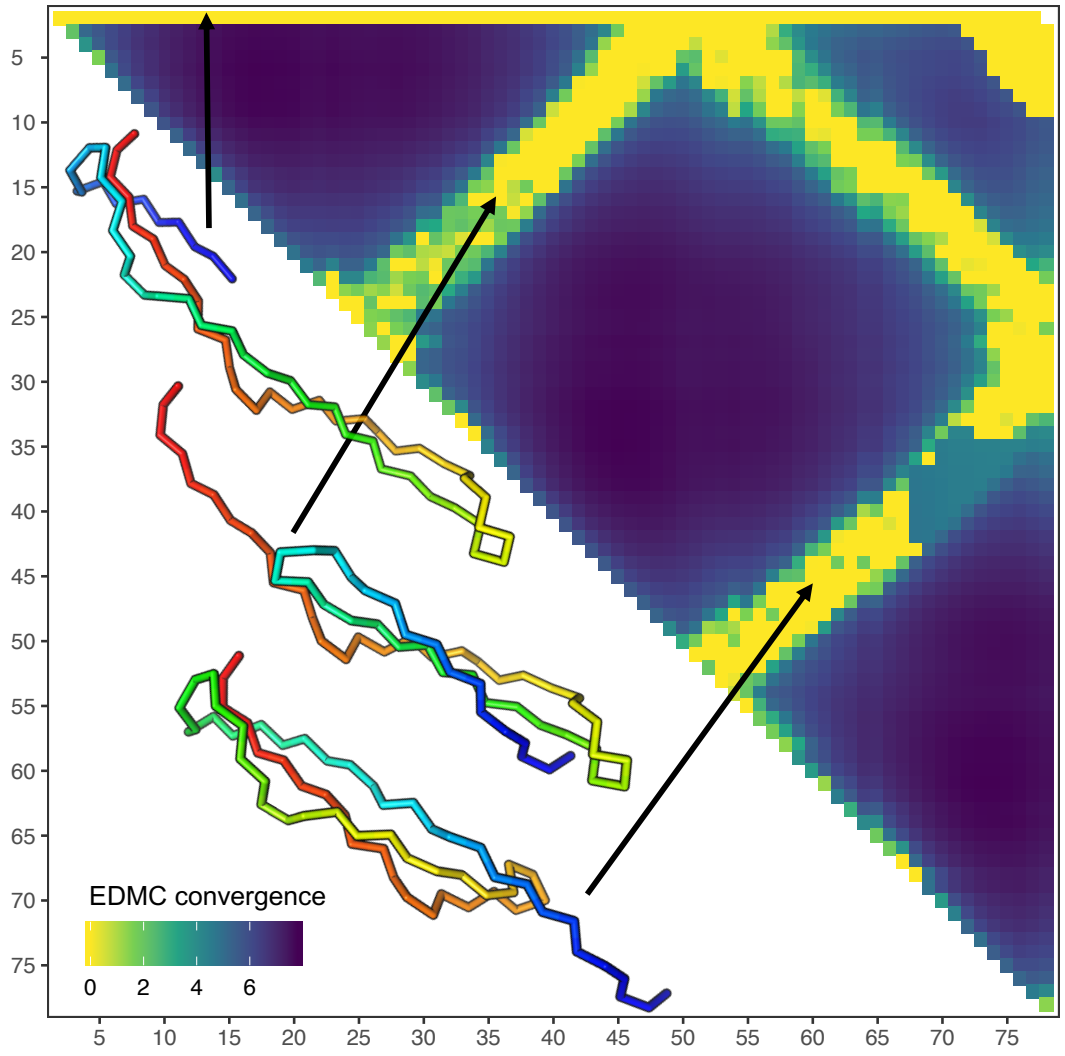

FIGURE 6.9 Modelling systematic structural deletions in the G5 domain of SasG using EDMs. All possible structural deletions in the SasG G5 domain structure (PDB:3TIP) were modelled with EDMs. Convergence values are shown in the upper diagonal part of the matrix: each entry corresponds to a deleted region in the domain from start (row number) to end (column number). Dark blue regions correspond to deletions with high model errors, suggesting they are geometrically invalid, and light yellow entries correspond to error-free models. Three valid models are shown and mapped to their corresponding deleted region: 1-15 (partial deletion of strand A), 16-35 (partial deletion of strands A and B), and 45-60 (partial deletion of strands B and C). The domain atrophy observed in the SasG E domain corresponds to deletion of the two regions 1-15 and 45-60, involving the N-terminal part of the single-layer A-B-C β-sheet.

domains also corresponds to regions around β-hairpin turns, D-E in Rib and A-B in SasG_E.

6.4 Discussion

In this final results chapter I have presented a new approach to model protein structures using Euclidean distance matrices (EDMs) that aims to be more intuitive, lightweight and faster than other traditional modelling alternatives based on molecular simulations at the task of modelling protein structural rearrangements. I have described how EDMs can be handled to suit different protein modelling needs, demonstrated their applications for structural rearrangements, and showed how EDM models can be used to analyse protein structure geometry and flexibility.

EDMs are well-suited to model protein structural rearrangements, where inter-residue distances at the domain core are conserved and only small regions in the native distance matrix are perturbed, and they can be naturally represented as matrix transformations. Furthermore, the convergence score of EDM completion algorithms can be used to estimate if a specific structural rearrangement is geometrically possible in 3D, providing insights into valid domain swap conformations and domain atrophy events in proteins.

I found that domain atrophy and swapping are highly constrained by geometry. Domain atrophies are predicted to be geometrically valid at β-hairpin turns, as observed in natural atrophy events in Rib and SasG_E domains. These result provides valuable insight into the structural mechanisms of domain atrophy and the evolution of domain structures. Furthermore, a high fraction of energetically favorable domain swap conformations were found to be geometrically impossible, suggesting that geometry is an equally important determinant of misfolding in tandem domain repeats.

EDMs have several limitations. Models are solely based on geometry, and it is not trivial how to incorporate other more complex energetic terms essential to protein folding and function. Additionally, there is a lack of specialised software libraries for EDM-based protein modelling, making it difficult to write efficient code for more sophisticated applications.

The current EDM modelling implementation does not scale well and proves

impractical for inputs over 600 atoms, mainly due to the matrix completion bottleneck, which means that most of the models are restricted to be coarse-grained (C_α or backbone only), even though the same approach described here can be used to generate full side-chain models. Adding residue side-chains to backbone models is, however, a common task in protein modelling and good tools already exist, for example Chimeras Dock Prep (Pettersen *et al.*, 2004).

The computational costs of EDM modelling could be reduced much further. On the software implementation side, matrix operations in `R` are two orders of magnitude slower than in other programming languages like `C++` and `Java`; using a faster implementation for EDM completion will have a big impact on the running time, but I am unaware of its availability. Another avenue to explore is the use of more sophisticated EDM completion techniques such as facial reduction (Drusvyatskiy *et al.*, 2017), which exploits rigid "cliques" of fully connected points to reduce the EDM size.

EDM modelling is well suited for tasks where a large proportion of atomic distances are known with high confidence, such as structural rearrangements. Other problems in protein modelling have similar conditions but have not been explored here, including modelling multidomain proteins by concatenating tandem or nested domains, docking protein subunits based on a subset of interface contacts, modelling allosteric conformational changes, and integrating distance constraints from experimental techniques such as NMR and cross-linking and computational distance predictions from sequence co-evolution inferences. EDMs are applicable to a broad range of molecular modelling problems and offer some advantages to traditional modelling techniques based on atomistic simulations; they should therefore be routinely considered for protein structure modelling and analysis.

6.5 Methods

6.5.1 Distance matrix index rearrangements

For an $n \cdot n$ Euclidean distance matrix A (Equation 6.2), I describe the index rearrangements used to model the three types of structural rearrangements discussed in this chapter.

Circular permutations

For a given native Euclidean distance matrix A with n rows and columns, the matrix A_{cp} of a circular permutation at "cut" position k corresponds to:

$$A_{cp} = \begin{bmatrix} A_{k+1..n,k+1..n} & A_{k+1..n,1..k} \\ A_{1..k,k+1..n} & A_{1..k,1..k} \end{bmatrix} \tag{6.3}$$

The terminal residues are connected in the circular permutation, so a terminal loop is created by setting their distances to other atoms as unknown. The length of the terminal loop can be extended by inserting additional columns and rows in the matrix.

Domain swapping

Starting from a Euclidean distance matrix A from a single native domain with n atoms, the matrix A_{swap} for a domain swap with "hinge loop" centred at position k corresponds to a duplicated matrix with $2n$ columns and rows and the following index rearrangements:

$$A_{swap} = \begin{bmatrix} A_{1..k,1..k} & \cdots & \cdots & A_{1..k,k+1..n} \\ \vdots & A_{k+1..n,k+1..n} & A_{k+1..n,1..k} & \vdots \\ \vdots & A_{1..k,k+1..n} & A_{1..k,1..k} & \vdots \\ A_{k+1..n,1..k} & \cdots & \cdots & A_{k+1..n,k+1..n} \end{bmatrix} \tag{6.4}$$

The entries of the matrix involving residues that form the "hinge loop" are however set as unknowns, allowing molecular flexibility to extend the loop and permit the interaction of the domains.

To model domain swap dimers, where interacting domains are from different subunits, atomic coordinates are assigned to different chains of the model. For tandem domain swaps, the domains forming the swapped conformation are part of the same protein chain, so an additional loop is required to connect the termini of the central domain at indices $(n, n+1)$, similar to circular permutations.

Domain atrophy

A structural deletion between positions k and l in a domain can be modelled by removing all columns and rows between indices k and l in the native Euclidean

TABLE 6.1 List of 20 ECOD domain structure representatives used to construct the lookup table of atomic distance bounds.

ECOD ID	PDB ID	PDB Range	UniProt ID	Architecture
e3s8iA1	3S8I	A:241-367	Q8WTU0	beta barrels
e1ublL1	1UBL	L:331-381	P21852	beta meanders
e1uaiA1	1UAI	A:2-224	Q9RB42	beta sandwiches
e1h6tA3	1H6T	A:78-240	P25147	beta duplicates or obligate multimers
e1rx0A1	1RX0	A:132-240	Q9UKU7	beta complex topology
e1gvdA1	1GVD	A:90-141	P06876	alpha arrays
e1t8kA1	1T8K	A:1-77	P0A6A8	alpha bundles
e1nkdA1	1NKD	A:1-59	P03051	alpha duplicates or obligate multimers
e1tu7A1	1TU7	A:78-208	P46427	alpha superhelices
e2gz4A1	2GZ4	A:6-205	Q7D028	alpha complex topology
e1kqfB3	1KQF	B:100-159	P0AAJ3	a+b two layers
e4iusA1	4IUS	A:1-106	D2PUG5	a+b three layers
e2qedA1	2QED	A:1-251	Q8ZRM2	a+b four layers
e3bqcA1	3BQC	A:3-330	P68400	a+b complex topology
e1g61A1	1G61	A:2003-2227	Q60357	a+b duplicates or obligate multimers
e5opqA2	5OPQ	A:316-625	G0L004	a/b barrels
e4om8A1	4OM8	A:1-181	Q988C8	a/b three-layered sandwiches
e4kb1A1	4KB1	A:8-215	P30014	mixed a+b and a/b
e1j0pA1	1J0P	A:0-107	P00132	few secondary structure elements
e2i5nH2	2I5N	H:1-36	P06008	extended segments

distance matrix A.

$$A_{atrophy} = \begin{bmatrix} A_{1..k,1..k} & A_{1..k,l..n} \\ A_{l..n,1..k} & A_{l..n,l..n} \end{bmatrix} \tag{6.5}$$

Residues k and l are connected through a peptide bond in the atrophied domain, so their distance are set according to protein geometry constraints.

6.5.2 Calculation of atomic distance bounds

I randomly selected one manual representative domain from each architecture type of the Evolutionary Classification of Protein Domains (ECOD) database (Cheng *et al.*, 2014). The set of 20 domain structures (Table 6.1) is of diverse length, secondary structure content and topology, and determined by high-resolution X-Ray crystallography (below 2Å).

I calculated over three million distances between backbone atoms (N, CA, CB, C, O) and grouped them by atom types and residue number separation along the peptide chain. Separations above five residues were grouped together into one category. For each of the 270 pairs of atom types and residue separation groups, I constructed a lookup table with the average value, the minimum value (lower bound) and the maximum value (upper bound). Outliers, defined as distances outside the 1/1,000 percentile, were discarded in the upper and lower bound calculations for robustness.

6.5.3 Implementation and availability

The protein modelling approach based on Euclidean distance matrices described in this chapter has been implemented in the `R` programming language and released as an open source repository that can be freely accessed under an MIT license at `https://github.com/lafita/protein-edm-demo`.

The repository contains scripts to model the structural rearrangement examples described in this chapter, the table of ECOD domain structures (Table 6.1), the lookup table of atomic distance bounds (data underlying Figure 6.4), and the running time estimates for ECOD domains (data underlying Figure 6.5).

The package `edmcr` (Rahman and Oldford, 2016) is used to access algorithms for EDM completion and multidimensional scaling, while protein structures are parsed and manipulated using the `bio3d` package (Grant *et al.*, 2006). The proposed implementation requires minimal dependencies and only a few lines of code, and runs on a single CPU.

Chapter 7

Conclusions

> *"To reveal the nature of the protein universe, we ask: How many protein sequences are there? How many sequences are novel vs. repetitious? How many sequences are characterized at structural and functional levels? Are sequences of prokaryotes, eukaryotes, and viruses different? Is the number of sequence families saturating or is it still expanding rapidly?"*
>
> – Levitt (2009): "Nature of the protein universe"

In this chapter, I summarise the research focus of this work and my aims and objectives, and discuss the key findings presented in previous chapters, along with the major challenges and limitations encountered. I also present remaining areas of work, discuss the most important implications deriving from this study for other areas of protein research, and make suggestions for future research avenues.

7.1 Overview

Domains are evolutionary units of proteins with independent functions and conserved globular structures (Hubbard *et al.*, 1999). Their diverse combinations in multidomain proteins — including domain duplications and other repetitions — are central to the evolution of proteins that carry out complex functions in cells (Chothia *et al.*, 2003; Levitt, 2009). The main focus of this work has been the study of tandem domain repeats, an understudied protein domain architecture consisting in sequential homologous domains.

There have been few comprehensive studies of tandem domain repeats in proteins, and they used known domain definitions from databases such as Pfam (Apic *et al.*, 2001; Björklund *et al.*, 2006). Over ten years after the publication of these studies, the exponential increase in protein sequence and structural data offered an opportunity to better understand the origin, evolution and role of tandem domain repeats in natural proteins. Advancements in bioinformatics methods for the detection of protein repeats also enabled unbiased approaches to detect tandem domain repeats and to discover new domain families not considered in previous studies.

Tandem domain repeats have been associated with several biological functions such as protein complex assembly, cell-adhesion and signaling; and important bacterial surface proteins involved in host invasion and biofilm formation are rich in large nearly identical tandem repeats. Experimental structures revealed that these repeating elements fold into stable globular domains (Gruszka *et al.*, 2012; Whelan *et al.*, 2019; Whelan *et al.*, 2020). This finding opened further research questions on the evolution of these proteins, observed to be highly variable and to function as rigid rods that form fibrils at the bacterial surface. It further motivated efforts to classify and characterise experimentally other similar tandem repeats in bacterial surface proteins.

Highly similar tandem domain repeats further challenge experimental observations by Wright *et al.* (2005) that highly similar adjacent domains in proteins have an increased rate of misfolding and aggregation. The sequences of adjacent homologous domains would therefore be expected to rapidly diversify or evolve misfolding-resistant mechanisms to avoid, or minimize to tolerable levels, their toxicity to cells and organisms caused by non-functional misfolded conformations and protein aggregates. Biochemical and computational studies suggested

that native-like interactions between adjacent domains are the main cause of misfolding in highly similar tandem domain repeats (M. B. Borgia *et al.*, 2011; Tian and Best, 2016), but studies are limited to small subsets of domains and general misfolding determinants in tandem domain repeats are still unknown.

The primary aim of this study was to understand the role of tandem domain repeats in natural proteins, and to discover unique sequence and structural properties potentially related to protein misfolding and aggregation. I planned to survey the prevalence and distribution of tandem domain repeats in natural proteins, investigate their emergence and evolution, and characterise their sequence and structural properties to understand misfolding determinants in tandem domain repeats. I also worked together with experimental collaborators to select target domains for further biochemical and structural studies.

This thesis is a computational study. I have used a wide range of bioinformatics methods, from protein homology and repeat detection to structural visualisation and comparison, to analyse large datasets of proteins and bacterial genomes. I have further developed and implemented two new computational methods to model the energetics and geometry of protein domains and their misfolded conformations, key tasks to approach some of the research problems in this study.

7.2 Key findings

The survey of tandem domain repeats presented in Chapter 2 showed that highly similar tandem domain repeats (above 90% sequence identity) are rare in natural proteins (up to 0.1% Eukaryotic and 0.04% Prokaryotic proteins), but repeats span a wide range of domain families in Pfam (14% of type domain families). These results were in line with previous findings by Apic *et al.* (2001) and Björklund *et al.* (2006). I observed that domain families with the highest prevalence of tandem domain repeats belong to a small subset of domain topologies, namely Immunoglobulin-like β-sandwich folds, Ubiquitin β-grasp folds and three helical bundles. Together with Alex Bateman, I further created 34 new families of tandem domain repeats in Pfam, improving their coverage by 8% (from 42% to 50%), and identified 68 domain families that form tandem domain repeats in bacterial surface proteins.

A large number of these tandem domain repeats are stalks in bacterial surface

proteins, and their function is to form rigid rods that project adhesive domains out of the bacterial surface to enable biofilm formation and host colonisation. Together with Jennifer Potts, we observed that these proteins exhibit length variation via repeat number changes, modulating their surface exposure, and we defined them as a new class of proteins named "Periscope proteins". In Chapter 3, I identified over 50 new groups of putative Periscope proteins with diverse domain compositions across a dataset of bacterial genomes. Although I did not find evidence of genomic repeat number variability in individual raw sequencing reads within genomic strains, I observed an evolutionary rapid and large variation of stalk domain repeats in homologous Periscope proteins of related bacterial strains.

In Chapters 2 and 3, I further found a widespread sequence composition bias in highly similar tandem domain repeats and across different Pfam families and Periscope proteins, with striking correlations between the bias and the sequence identity of adjacent domains in some families. I found that the sequence bias is stronger in bacterial surface proteins, and although it varies across proteins, certain amino acids such as Thr, Pro, Gly, and Ala are frequently enriched (although they are also depleted in other proteins). I further observed that, in Periscope proteins, the amino acid bias is specific to the stalk domain repeat regions.

Tandem domain repeat structures analysed in Chapter 4 have revealed a high structural malleability, with rare evolutionary recent cases of atrophy and elaboration in domains such as Rib and SasG_E, and suggest the action of unusual evolutionary and selection pressures in tandem domain repeats. Other domains, such as SHIRT and domains in the MBG superfamily, show atypical folds only remotely related to other known β-grasp and β-sandwich topologies, with little structural similarity to other domains in the PDB.

In Chapter 5, I found that domain families commonly found in tandem repeats are predicted to be more resistant to a certain type of protein misfolding, known as domain swapping. I identified three potential misfolding determinants: domain topology (related to the orientation and interaction of the N- and C-termini), the domain length and the inter-domain linker. Certain domain topologies — such as ones commonly found in tandem repeats: β-sandwich, β-grasp and three-helix bundle folds — are more resistant to domain swap misfolding, but there is a large variability in the predicted misfolding propensity of domains within domain families, suggesting that changes in the sequence and structure of these domains can

have large effects on their misfolding propensity. One of the factors I identified as important is the length of the domain: shorter domains were predicted to be more resistant to misfolding, likely due to shorter and more rigid loops.

I finally showed how structural rearrangements such as domain atrophy and swapping can be naturally represented as Euclidean distance matrix (EDM) transformations, and used to model protein conformations in 3D. In Chapter 6, I use this EDM modelling approach to predict geometrically valid structural rearrangements in protein domains, providing insights into valid domain swap conformations and domain atrophy events in proteins. I found that domain atrophy and swapping are highly constrained by geometry: valid structural deletions are predicted to be around β-hairpin turns, as observed in the natural domain atrophy events of Rib and SasG_E, and some energetically favorable domain swap positions are predicted to be invalid protein conformations. These results are important steps towards understanding domain structure evolution and misfolding at the molecular level.

7.3 Challenges and limitations

Throughout this study, I encountered several challenges and limitations. Some of the limitations are related to computational methods used and developed, while others involve gaps in the amount and types of available data. Despite our efforts, some of our initial research questions could not be fully addressed and remain partially open.

7.3.1 Methodological limitations

The computational detection of tandem domain repeats proved to be challenging. *De novo* approaches make use of heuristics to reduce the computational complexity of the problem and require custom parameters (Lim *et al.*, 2013), which cause instabilities in the number and boundary of repeats and only permit the detection of highly similar repeats (at 80-90% sequence identity). As a result, only small subsets of highly similar domain-size tandem repeats can be detected, and these do not always correspond to globular domains. I aimed to tackle these limitations using a second tandem repeat detection approach based on pre-existing Pfam domain family models, which is more robust in the number and boundary of repeats

and it extends the search to more remote domain repeat homologs (below 20% sequence identity), but it is limited to known globular domain families only. By combining these two approaches, I improved the coverage of tandem domain repeats in Pfam, but it remains lower (around 50%) than the Pfam sequence coverage above 70% (El-Gebali *et al.*, 2019).

Contrary to tandem repeat detection methods, there were no computational tools openly available to predict and model domain swapping and circular permutations in protein domains, despite several previous studies had described such methods (Paszkiewicz *et al.*, 2006; Ding *et al.*, 2006; Lo *et al.*, 2012b; Tian and Best, 2016). I decided to start a collaboration with Pengfei Tian and Robert Best (National Institutes of Health, USA) to implement an improved and automated version of their alchemical method that could be scaled to analyse large datasets of protein domains, named TADOSS. I released the source code openly, with documentation on how to install and use the method, hoping it would be useful to other researchers interested in predicting circular permutations and domain swaps (Lafita *et al.*, 2018).

The TADOSS method predicts stable domain swapping positions in a domain, but is solely based on the energetics of the native structure and makes several assumptions. It only considers the ideal case of two identical domains in tandem, but a more realistic scenario would take into consideration some amount of sequence diversity, as the majority of tandem domain repeats found in nature are not identical. Furthermore, TADOSS only takes into account the energetics of the native structure, assuming that most native contacts remain unaltered and no new interactions occur, but other features might also be important for the formation of tandem domain swaps, such as additional stabilizing contacts formed at the hinge loop and the geometry of the protein backbone.

Structural models of domain swap predictions are essential to take into account the protein geometry and predict hinge loop contacts. The modelling approach based on Euclidean distance matrices (EDMs) that I presented is solely based on protein geometry, and although it can be used to model domain swap conformations and evaluate their geometrical feasibility, it does not consider energetic constraints essential to protein folding and stability. In addition, the lack of specialised software libraries for protein EDM modelling makes it difficult to write efficient code. The current EDM modelling implementation does not scale

well and proves impractical for models over 600 atoms, which means that models are restricted to coarse-grained representations (C_α or backbone only); although these were sufficient for the protein modelling applications explored in this study.

7.3.2 Data availability

The vast majority of genomes are sequenced using short-read technologies such as Illumina. In the study of Periscope proteins, I used a specialised dataset of bacterial genomes sequenced with long-read PacBio technology (named NCTC3000 genomes) in order to obtain reliable estimates of the repeat numbers in assembled genes and to be able to analyse raw sequencing reads covering entire genes (over 3,000 bases). A major limitation of the NCTC3000 dataset was the small number of bacterial strains and species available (734 strains from 207 species at the time of this study), which hindered the discovery of a larger number of putative Periscope proteins and the study of domain repeat number variations across strains. Another limitation was the sequencing depth of the genomes, which proved to be insufficient to observe repeat number variations within a single genome *in silico*. Furthermore, the repeat detection from raw PacBio sequencing reads posed many challenges due to their low quality (reads were only 80% sequence identity to assembled genes) and instability, involving rare large insertions in the repeating regions of the genes.

Despite our structural determination efforts and the progress from our collaborators, several tandem domain repeats remain uncharacterised. State-of-the-art structure predictions provided by the Baker lab have greatly increased the structural coverage for Pfam families, for example in domain families from the MBG clan, but some families could not be modelled due to the limited number of sequences (such as SSURE) or proved difficult to model due to their rare sequence composition and large structural changes (such as the likely wrong YDG model). In addition, the number of experimental structures for constructs of tandem domains is still very limited, hindering important analyses of inter-domain linkers, domain interactions and orientations.

The amount and diversity of folding experiments on tandem domain repeats is scarce — experiments are low-throughput and have been done on a limited set of protein domains, namely titin I27 (Wright *et al.*, 2005; M. B. Borgia *et al.*, 2011) and spectrin (Batey *et al.*, 2008) — posing a challenge for theoretical and

computational studies such as this one. Despite many types of molecular misfolding have been experimentally observed, only certain types of idealised misfolded conformations, such as tandem domain swaps considered in this study, can be addressed computationally. A further knowledge gap is the lack of an experimental structure of a tandem domain swap conformation, which would be very helpful to set a precedent for future studies and to test predictions and models. There are sensible reasons for its absence: the low prevalence of tandem identical domain repeats in natural proteins, the selective reduction of repeats in crystallization constructs, and the transient nature and heterogeneity of tandem domain swap conformations. The similarity of tandem domain swaps to other better characterised protein variations, namely circular permutations and domain swap oligomers, has however permitted us to test tandem domain swap predictions and models to a certain degree.

7.3.3 Open questions

The causes and consequences of the amino acid composition bias observed in highly similar tandem domain repeats are still poorly understood. Although I have identified the enrichment for a subset of common amino acids to be the cause of the bias, other amino acids are enriched in some domains but not in others. I explored several general properties of amino acids, such as side-chain entropy, hydrophobicity and charge, but none of them generally correlates with the type of amino acid composition bias observed, and direct links to protein aggregation or misfolding resistance mechanisms of the bias could not be found.

Another open area of research is the molecular mechanisms and frequency by which highly similar tandem domain repeat regions in Periscope proteins expand and contract. I have suggested DNA recombination driven by nearly identical repeats as a possible molecular mechanism, since domain repeat changes occur among domain repeats with the highest similarity in central regions of Periscope genes, but more experimental evidence is needed. I observed dramatic changes in the repeat number between similar bacterial strains, but I could not observe variability within strains at sequencing depths of around 100 reads, suggesting that these molecular mechanisms occur with frequencies lower than 1/100.

The structural malleability observed in tandem domain repeats leaves many open questions about the impact of structural changes, such as domain atrophy,

to the domain stability. Given the magnitude of the structural changes, and other compensatory mutations such helical elaborations, it is infeasible to predict them computationally, so custom experiments would need to be designed to specifically test them.

7.4 Implications and future perspectives

This thesis constitutes the most comprehensive study of tandem domain repeats in proteins to date, and the first one to focus on nearly identical repeats. Understanding the roles of tandem domain repeats in natural proteins has important implications for other areas of biological research. For example, in microbiology tandem domain repeats are key for bacterial pathogenicity in surface proteins implicated in cell-adhesion and biofilm formation. The improvements in classification of tandem domain repeats in bacterial surface proteins enable the development of computational methods to systematically predict pathogenic strain phenotypes from their genomes, a project already started by Monzon *et al.* (2020).

This study has also contributed to our understanding of the origin and evolution of globular domain structures, and has the potential to improve structure classification efforts by providing links between seemingly unrelated domains. I have suggested several domain targets for further experimental determination and started some through our collaborators, and I expect other research groups to attempt to discover further cases of structural malleability and evolution in these domains and study their implications for protein stability and function.

Our work has further implications to understand protein misfolding mechanisms and for the engineering and design of multidomain proteins. Knowledge of multidomain protein folding, and the impact of tandem domain repeats, not only improves our understanding of protein misfolding diseases, but also improves our ability to produce better protein therapeutics. Multidomain proteins will become more important with increasing complexity of protein designs, for example multifunctional and multispecific proteins that bind more than one target, usually an effector and a target protein (Deshaies, 2020). Furthermore, constructs of tandem domain repeats can be used to increase binding avidity and are desirable for design simplicity and convenience.

7.4.1 Experimental perspectives

Several new experimental techniques offer unique opportunities to study tandem domain repeats. Cryo-electron tomography has emerged as an exceptional tool to study biological macromolecules *in situ* within cells (Schaffer *et al.*, 2019), and several experimental groups have already used it to observe thin bacterial surface fibrils formed by tandem domain repeats such as CdrA (Melia *et al.*, 2021). This technique can already be used to tackle research questions such as the relation between bacterial cell envelope thickness and number of domain repeats in Periscope proteins and the protein density at the bacterial surface, both hard to predict computationally from the genomic sequence alone. If near-atomic resolutions can be reached, cryo-electron tomography has the potential to reveal the role of tandem domain repeats in creating specific molecular interactions and potential protein misfolding events in cellular conditions; I expect this technique to be more routinely used thanks to its many upsides.

Techniques to study proteins in high-throughput are becoming more accessible, and will be needed to bridge the gap between individual experimental data and large-scale computational analyses observed in this study. Protein structure determination services such as X-Ray crystallography facilities in Synchrotrons are already highly automated and offer pipelines to solve tens of structures concurrently, although some challenges still remain as evidenced in this project. Modern experimental setups allow high-throughput protein stability measurements by integrating protein expression, purification and stability testing in parallel (Perez-Riba and Itzhaki, 2017). These techniques will be useful to tackle open questions about what are the most malleable parts of tandem domain repeats and investigate the stability and misfolding propensity of adjacent domains with high sequence similarity for a wider range of domain folds, providing larger experimental datasets to test computational predictions. Studies of systematic protein stability measurements have proved extremely useful in the past, such as work on circular permutations of a DHFR domain by Iwakura *et al.* (2000).

Finally, tailored deep genome sequencing efforts with long-read techniques such as PacBio or Nanopore are needed in order to increase the read coverage of repetitive genes to sufficient levels in order to observe changes in the repeat number within strains. These techniques still suffer from low quality and unstable raw reads, complicating downstream computational analyses, but have the potential

to unravel the impact of environmental effects and selection pressures (such as high temperature or hostile conditions) on the number of domain repeats and the frequency of mutations in bacterial surface proteins.

7.4.2 Computational perspectives

The exponential increase in biological sequence databases is breathtaking. Protein sequences from metagenomics already reach the billions (Mitchell *et al.*, 2018), with broad impact for protein classification, the development of prediction methods and evolutionary studies. These large amounts of sequence data offer great potential for research in computational biology, but require new and more efficient computational methods and approaches.

Accurate protein structure predictions have a huge potential to revolutionise computational studies like this one, shifting discovery and evolutionary analyses from the sequence space to the structural domain, and offering additional molecular and mechanistic insights. I showed examples of this trend in the analysis of tandem domain repeat structures, where models by trRosetta (J. Yang *et al.*, 2020) have been key to interpret the structural differences observed among the domains in the MBG superfamily, but more sophisticated large-scale analyses using protein models might be feasible in the near future.

7.5 Concluding remarks

This study has focused on a very special subset of proteins, tandem domain repeats, which are rare in natural proteins but have proven to be exceptional in many ways: they form some of the longest proteins ever known in Eukaryotes and Bacteria, they have extremely biased sequences that mislead protein homology search and prediction methods, and they evolve rapidly, both at the protein level by changing the number of repeats and at the structural level through rare domain atrophy and elaboration mutations. Tandem domain repeats have challenged several of our assumptions about protein folding and evolution, and they have taught us fundamental principles that help us better understand the nature of proteins.

Appendix A

Tandem domain repeats in Pfam: additional plots

This appendix includes additional plots of sequence properties calculated for a few Pfam domain families with a high prevalence of tandem domain repeats. All figures in this appendix have the same subplot arrangement; the following figure legend applied to all of them:

From left to right, top to bottom. **First row:** comparison of the distribution of the sequence bias (measured as Shannon entropy), average residue side-chain entropy, and domain length for isolated, tandem domains with low sequence identity (<70%) and tandem domains with high identity >70% (hitrds). **Second row:** scatter plots of sequence bias, average residue side-chain entropy and domain length against the sequence identity of adjacent domains, split by organism Superkingdom. **Third row:** fraction of gaps and average insertions for each position in the HMM profile model of the family. **Last row**: distribution of the sequence identity between adjacent, non-adjacent and non-tandem domains in the same protein (left); scatter plot comparison of sequence identity between adjacent and non-adjacent domains (middle); and linker distance between adjacent domains as a function of their sequence identity (right). The pairwise sequence identity between domains is calculated from the Pfam sequence alignment (omitting gaps).

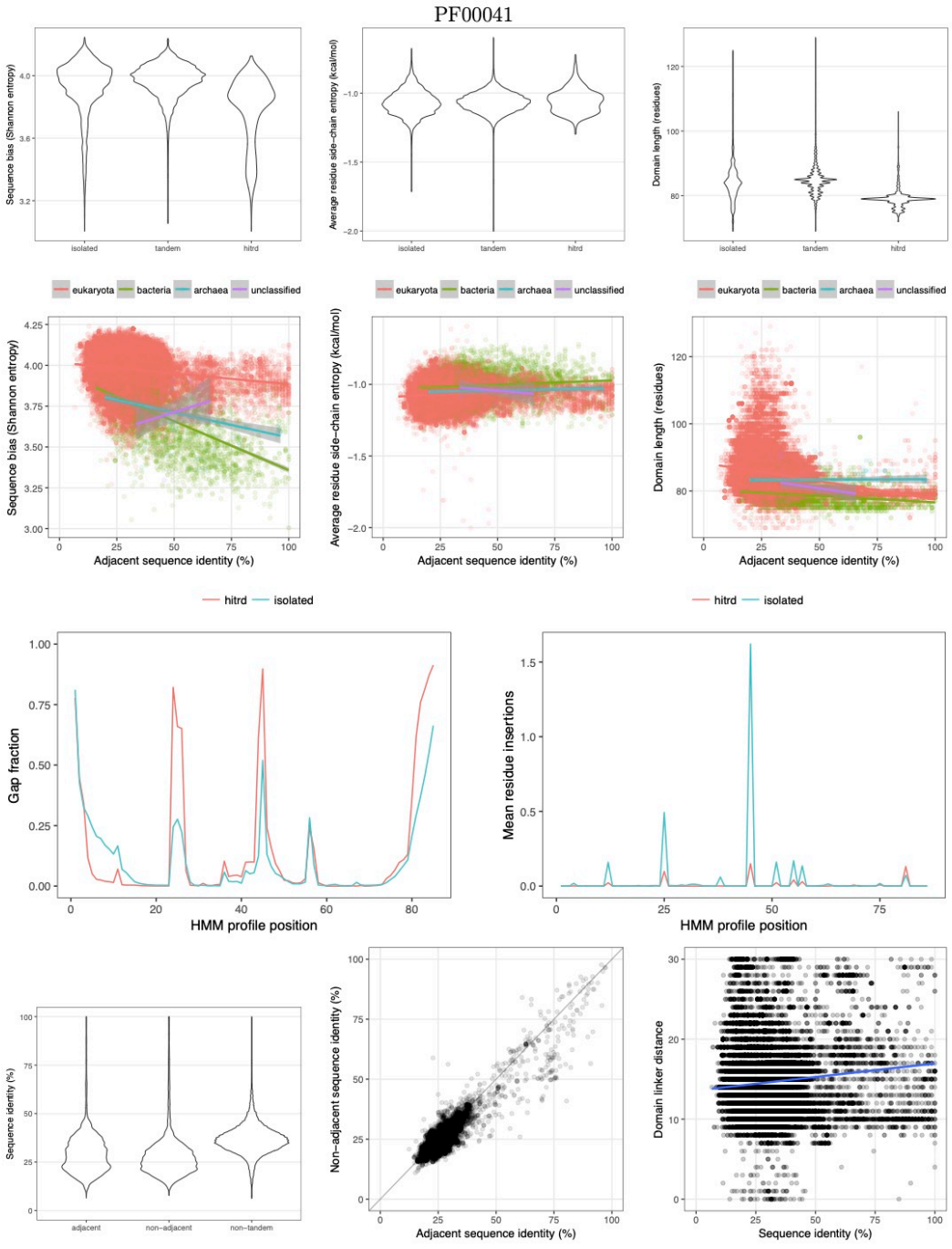

FIGURE A.1 Properties of tandem domain repeats in the fn3 domain family (Pfam:PF00041). For figure details see cover page of this Appendix A.

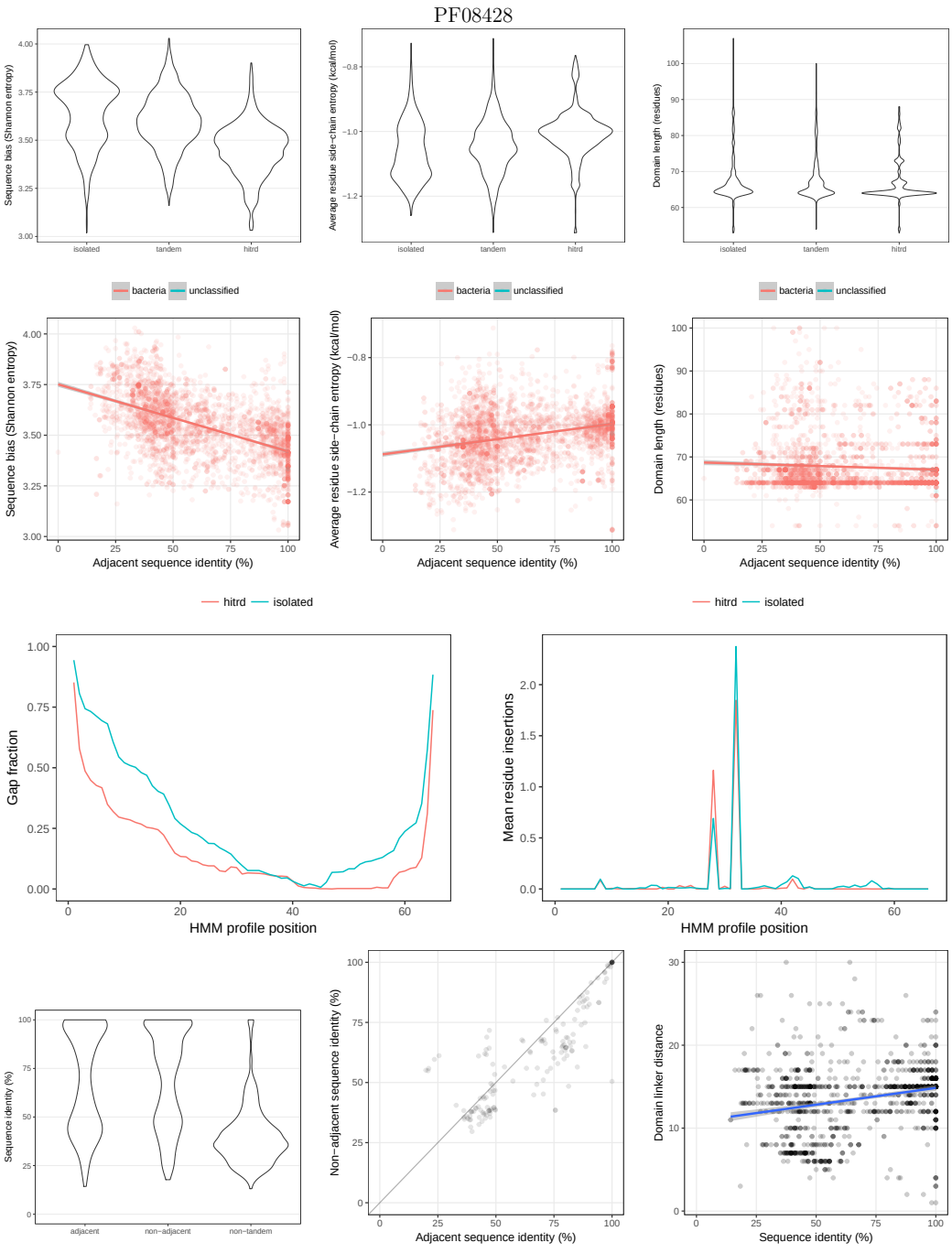

FIGURE A.2 Properties of tandem domain repeats in the Rib domain family (Pfam:PF08428). For figure details see cover page of this Appendix A.

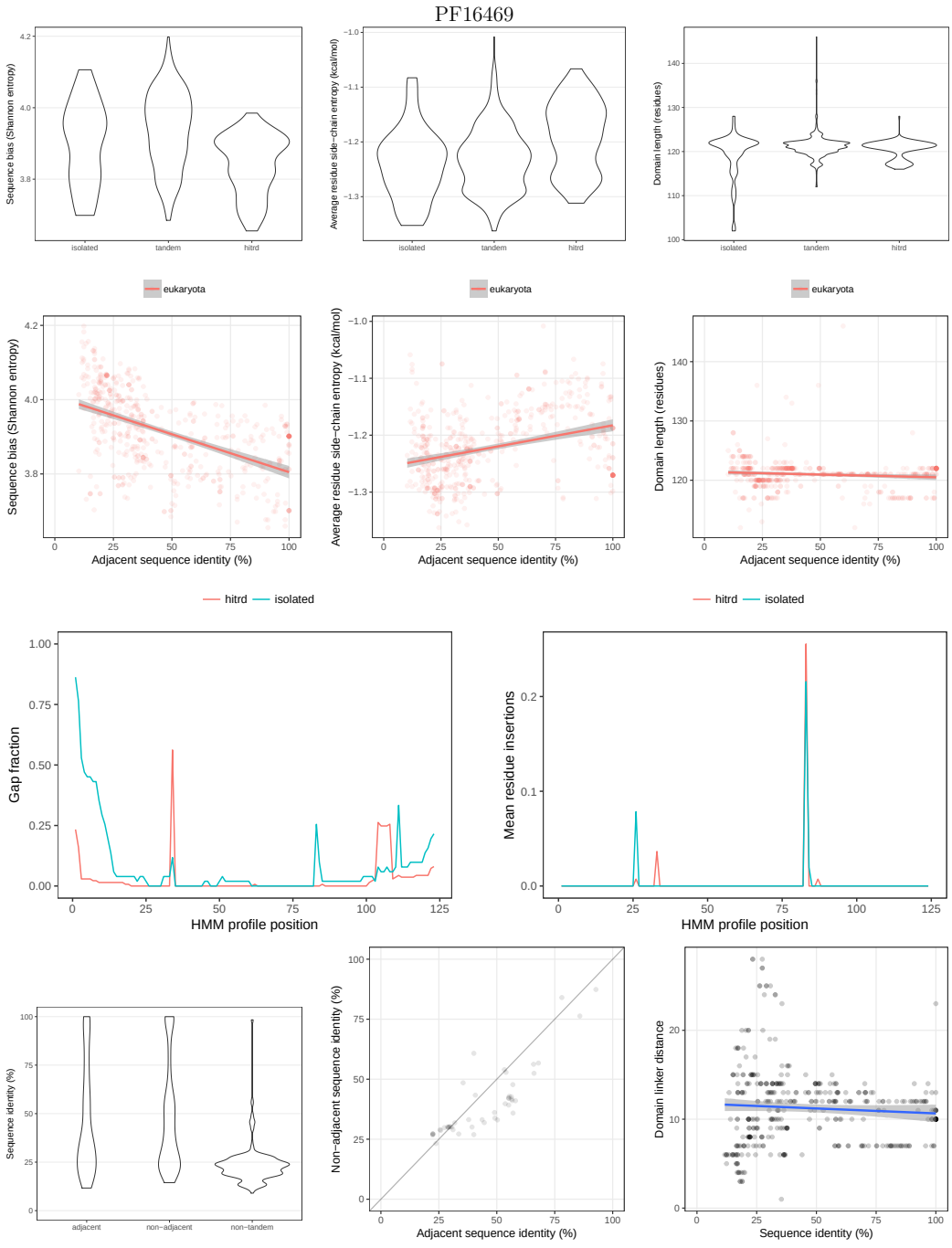

FIGURE A.3 Properties of tandem domain repeats in the Nematode polyprotein allergen ABA-1 (NPA) domain family (Pfam:PF16469). For figure details see cover page of this Appendix A.

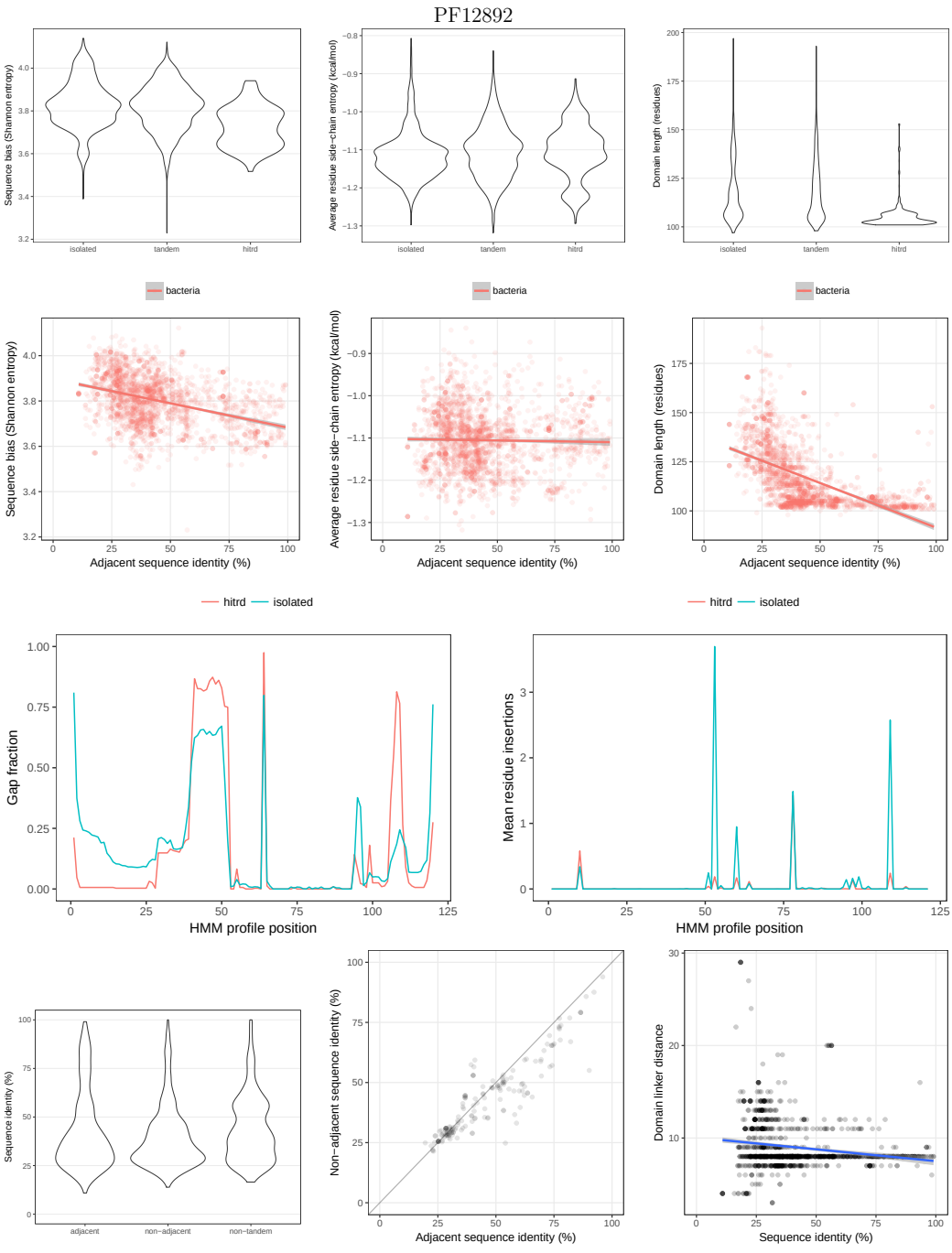

FIGURE A.4 Properties of tandem domain repeats in the FctA domain family (Pfam:PF12892). For figure details see cover page of this Appendix A.

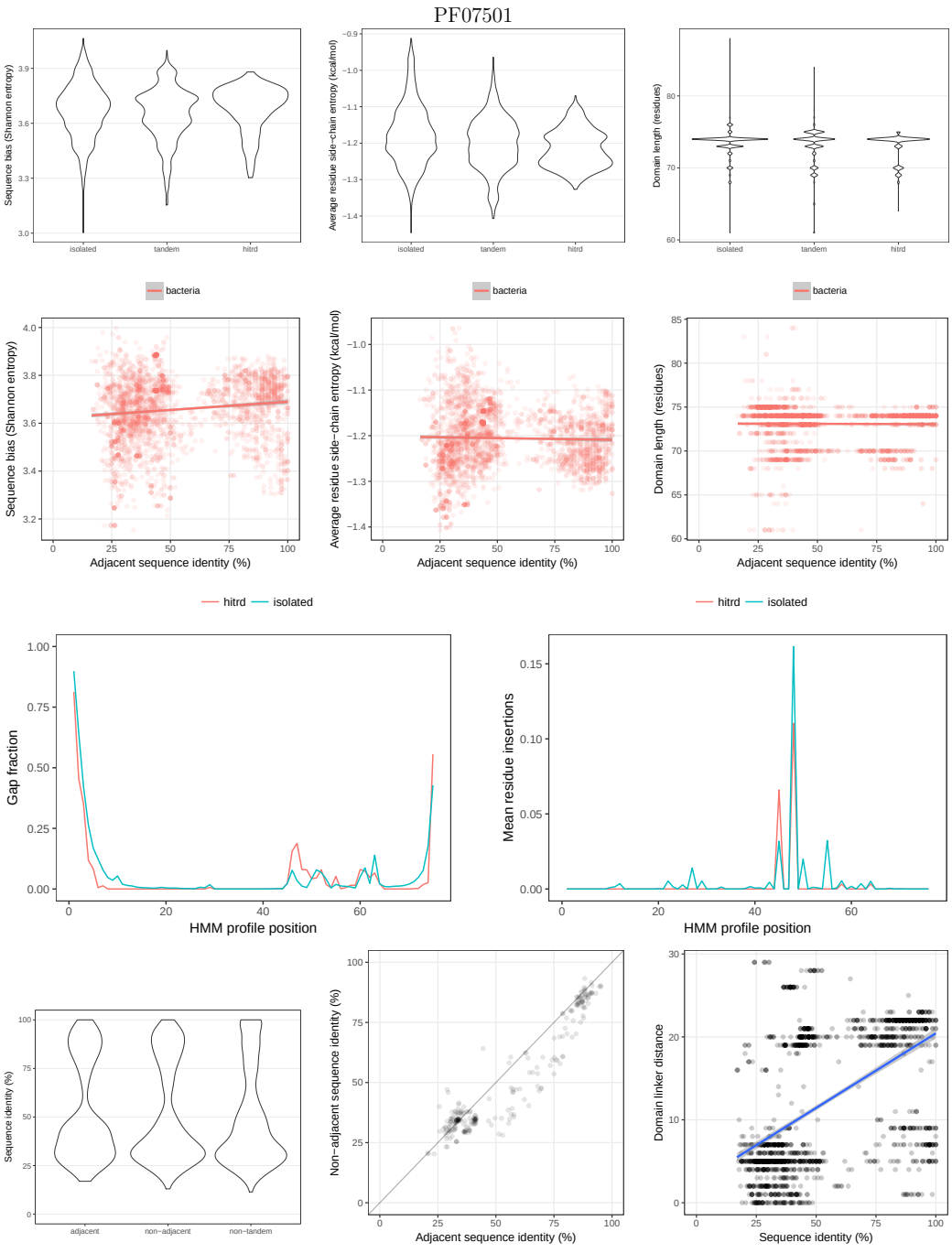

FIGURE A.5 Properties of tandem domain repeats in the G5 domain family (Pfam:PF07501). For figure details see cover page of this Appendix A.

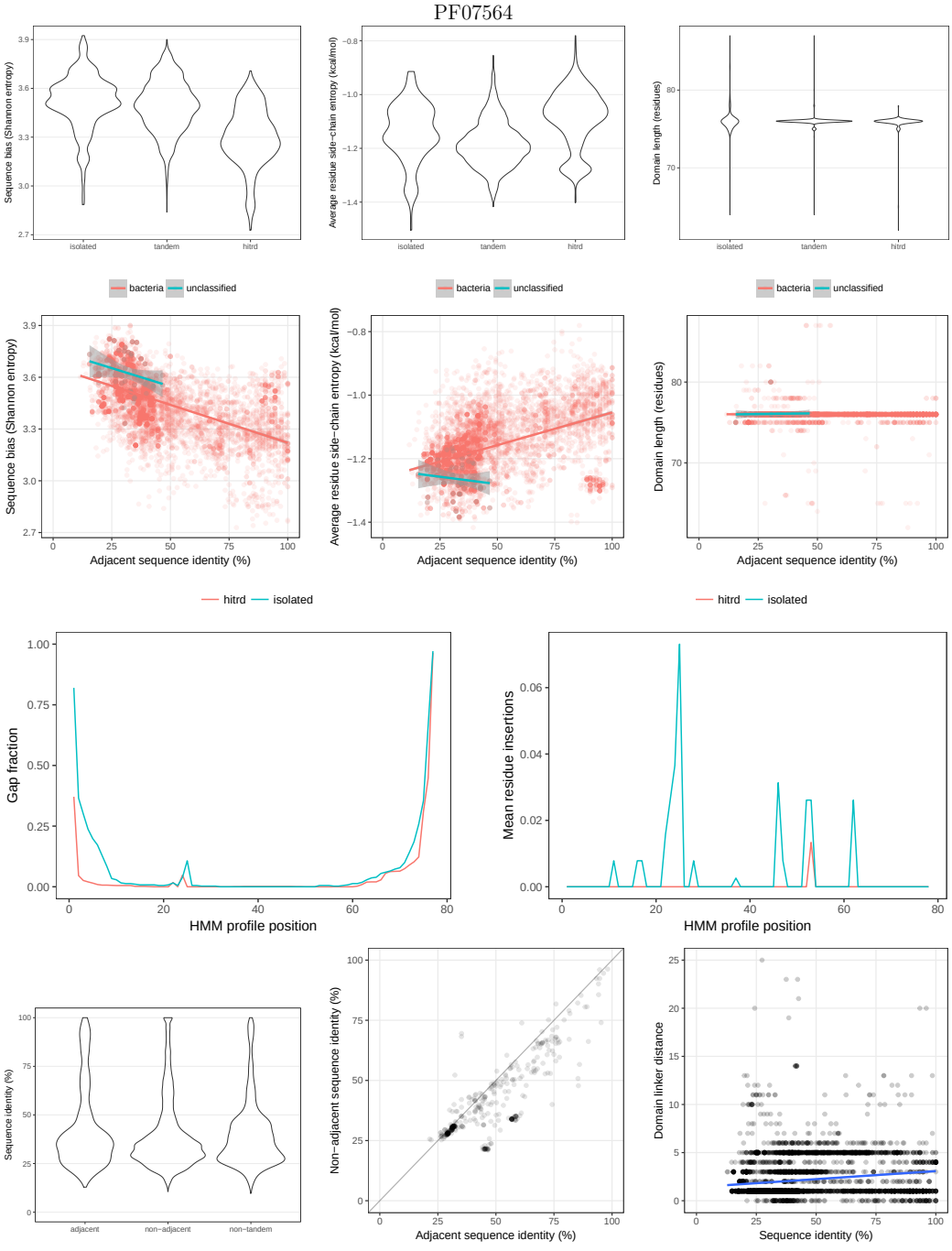

FIGURE A.6 Properties of tandem domain repeats in the DUF1542 domain family (Pfam:PF07564). For figure details see cover page of this Appendix A.

Appendix B

Periscope proteins: additional plots

This appendix includes additional plots on the analysis of Periscope proteins.

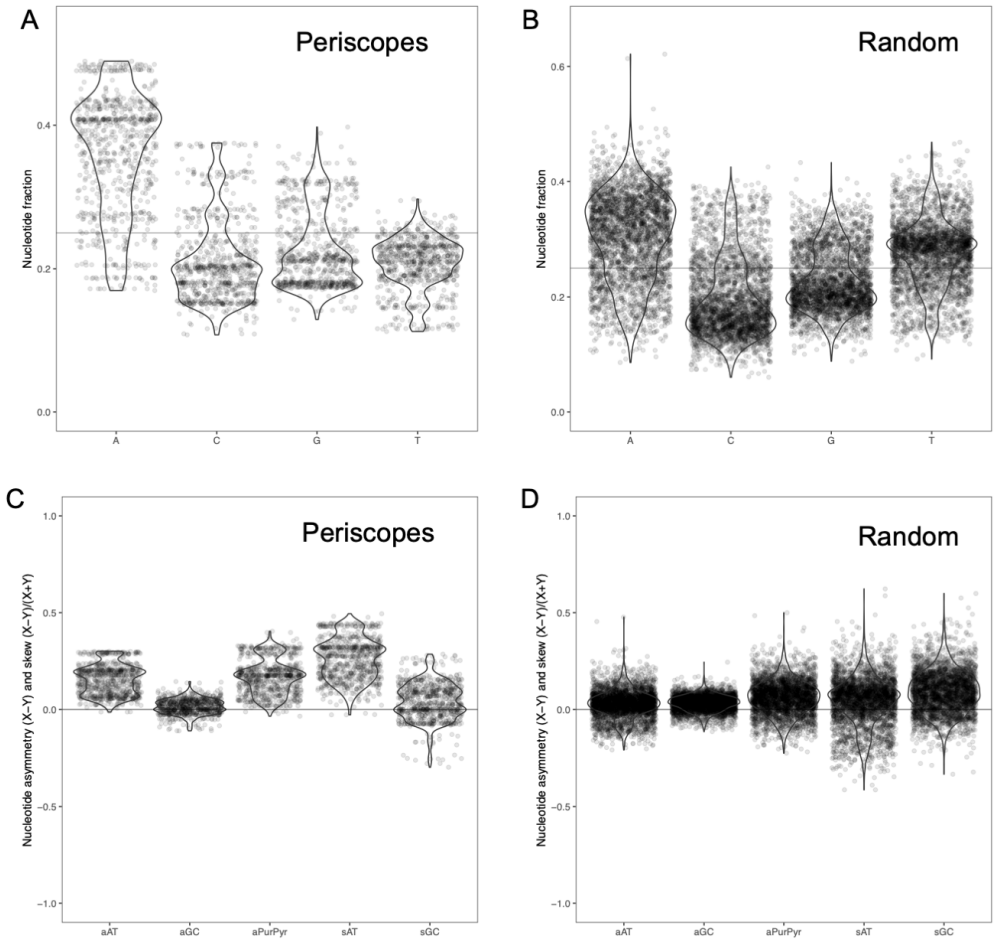

FIGURE B.1 Nucleotide composition skew of Periscope genes Violin plots of nucleotide composition in all putative Periscope genes (A) and an equivalent balanced subset of random genes from the same NCTC3000 genomes (B), and associated skew metrics (C, D): aAT (AT asymmetry), aGC (GC asymmetry), aPurPyr (Pyridine-Purimidine asymmetry), sAT (AT skew), and sGC (GC skew).

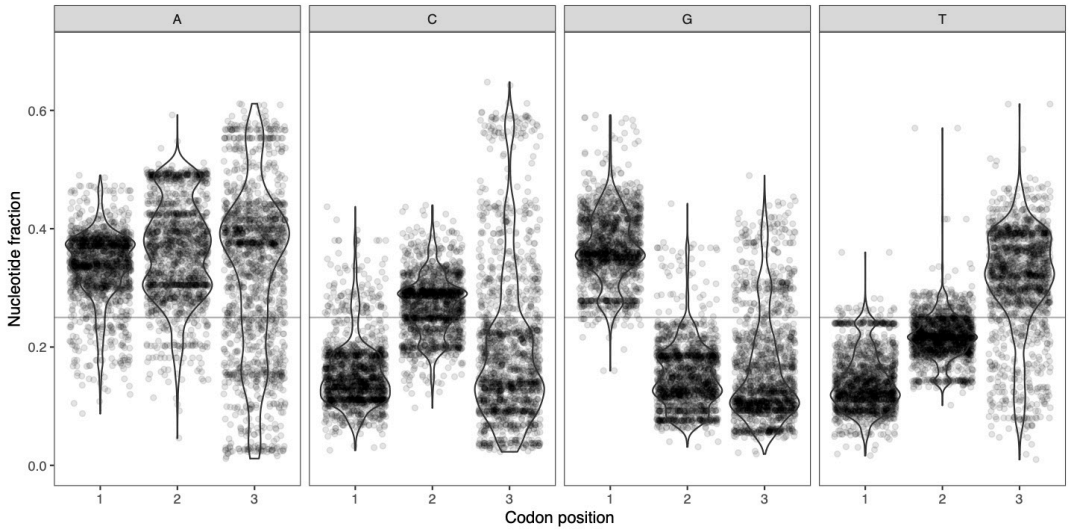

FIGURE B.2 Nucleotide composition in Periscope stalk repeats by codon position. Violin plots of nucleotide composition in sequences of stalk domain repeats of Periscope genes, split by codon position and nucleotide.

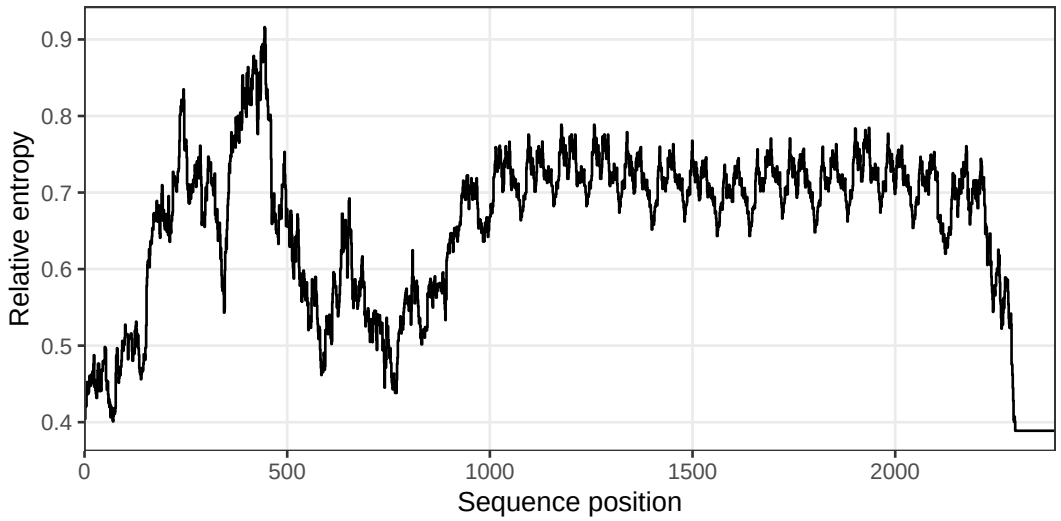

FIGURE B.3 Sequence bias profile of CdrA. Sequence bias along the protein measured as the relative entropy, using a rolling average with a window size of 100 residues. Higher entropy values represent more biased compositions.

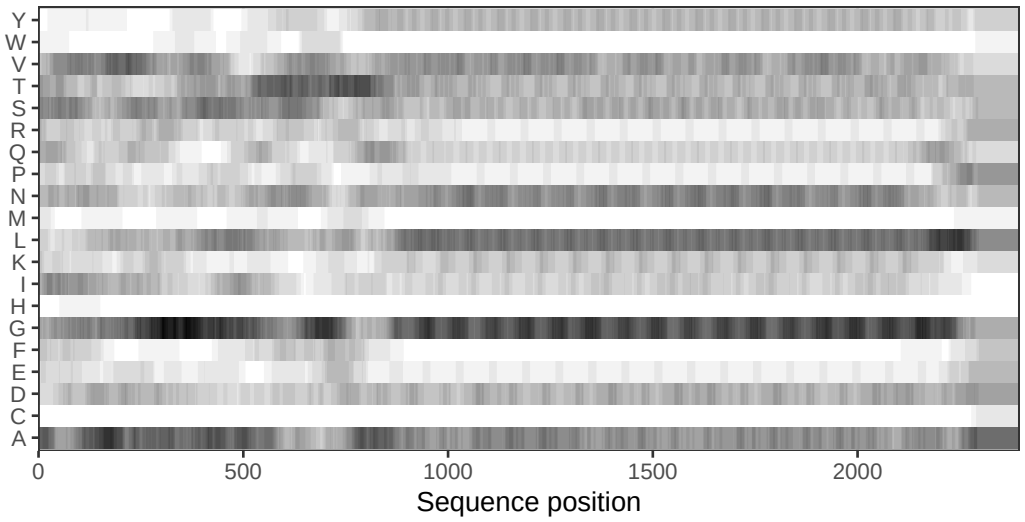

FIGURE B.4 Amino acid composition profile of CdrA. Amino acid composition along the protein measured as the fraction of each amino acid, using a rolling average with a window size of 100 residues, and colored in a grey scale (white to black) for fraction values in the 0-1 range.

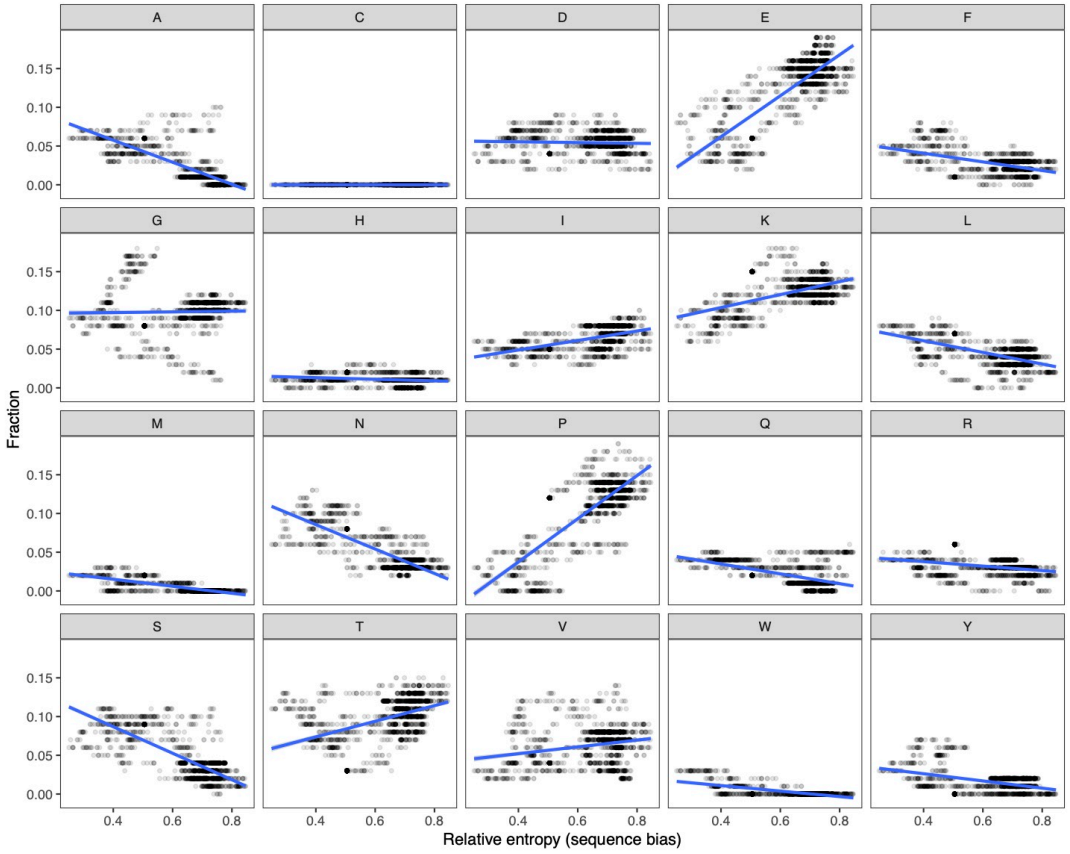

FIGURE B.5 Sequence bias and amino acid correlations in SasG. Analysis of the correlation between the protein sequence bias (x-axis) and the amino acid fraction (y-axis), split for each of the 20 amino acids. Linear fit for each plot is shown as a blue line.

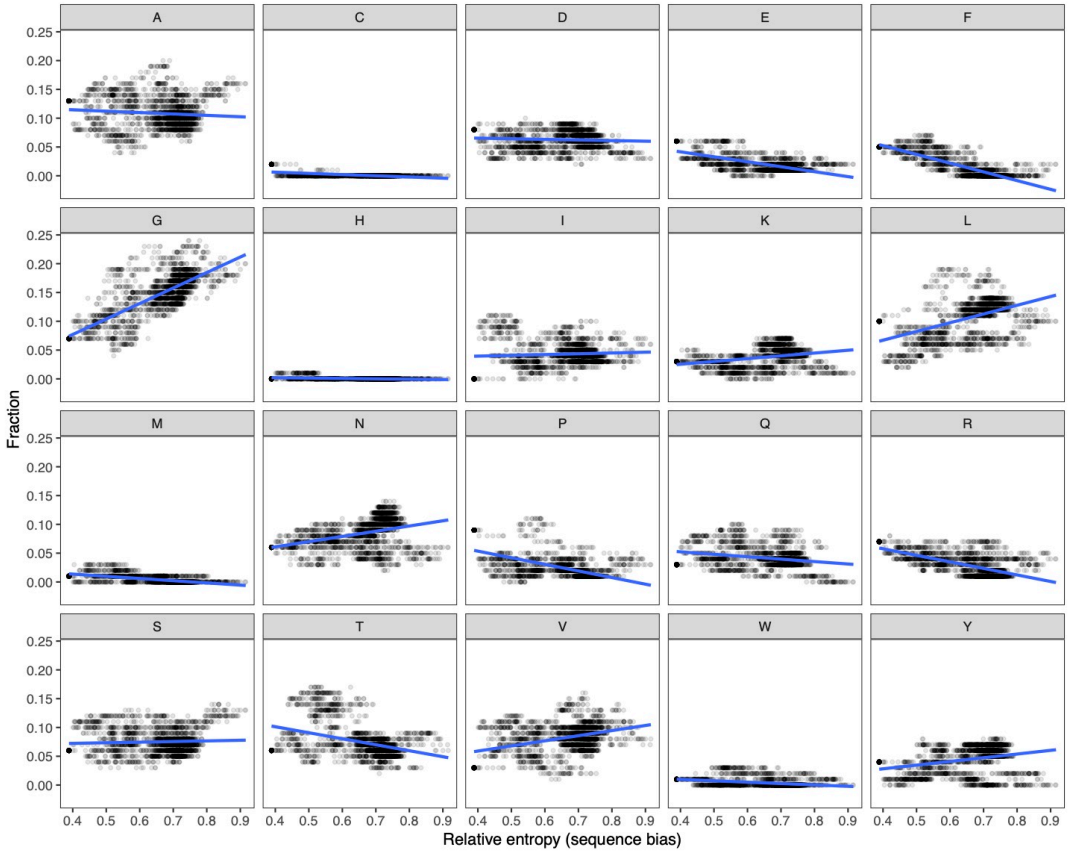

FIGURE B.6 Sequence bias and amino acid correlations in CdrA. Analysis of the correlation between the protein sequence bias (x-axis) and the amino acid fraction (y-axis), split for each of the 20 amino acids. Linear fit for each plot is shown as a blue line.

Appendix C

TADOSS: alchemical free energy model

A version of this energetic model was included as a Supplementary material in the TADOSS article (Lafita *et al.*, 2018), and it can also be found in the following GitHub repository: https://github.com/lafita/tadoss

The relative stability of a native domain (WT) with respect to its domain swapped conformation (SW) is defined as a difference in free energy:

$$\Delta\Delta G = \Delta G_{WT} - \Delta G_{SW} \tag{C.1}$$

$\Delta\Delta G$ is negative if the native state is more stable than the swapped state.

Each free energy difference (ΔG) represents the change between folded and unfolded absolute free energies:

$$\Delta G = G_F - G_U \tag{C.2}$$

Since native and the swapped conformations have the same amino acid sequences, their unfolded states have the same absolute free energy (G_U). Hence, their relative stability only depends on the absolute free energy of their folded states:

$$\Delta\Delta G = G_{WT} - G_{SW} \tag{C.3}$$

The free energy difference can also be split into the change in energetics (E) and entropy (S) between the two states, where T is the temperature:

$$\Delta G = \Delta E - T\Delta S \tag{C.4}$$

Since the majority of native residue contacts are conserved in the swapped conformation, their energetic and entropic components remain unaltered and their contribution to the free energy difference ($\Delta\Delta G$) will be negligible. The $\Delta\Delta G$ can therefore be estimated (as $\Delta\Delta G_{alchemical}$) using the sum of energetic contributions ϵ from each disrupted (broken) native contact C and the total gain of entropy from the native conformation, calculated as the entropic gain δs from each disrupted (unfolded) residue R_U in the domain swap:

$$\Delta\Delta E_{alchemical} = \sum_{c}^{C} \epsilon_c \tag{C.5}$$

$$\Delta\Delta S_{alchemical} = R_U\ (-\delta s) \tag{C.6}$$

The energetic component will have a negative contribution to the total free energy (contacts are formed going from the swap to the native conformation), while the entropic component will have a positive contribution (unfolded residues in the swap become folded in the native structure). The alchemical $\Delta\Delta G$ can be expressed as:

$$\Delta\Delta G_{alchemical} = \sum_{c}^{C} \epsilon_c - T \cdot R_U\ (-\delta s) \tag{C.7}$$

We can further distinguish two sets of native contacts that will be disrupted in the domain swapped state: 1) contacts at the termini, disrupted when residues are peeled off the the N- and C-terminal when they are covalently joined, and 2) contacts at the hinge loop, disrupted when a region of the domain is extended to form a linker between the two domains. For consistency with the original method description, the alchemical free energy contributions are defined as following:

$$\Delta\Delta G_{alchemical} = \Delta G_J + \Delta G_C \tag{C.8}$$

The topological requirement for joining the termini is that the linear distance between residues at position i and j is shorter than the effective length contributed by the peeled off residues:

$$d(i,j) < (i + L - j - M - L_l)\ r_0 \tag{C.9}$$

Where L is the length of the protein (number of residues), L_l is the inter-domain linker length and M is a penalty for the need of a loop when the termini point in opposite directions, defined as:

$$M = 6\ \sin(\frac{\theta}{2}) \tag{C.10}$$

In the above expression, θ is the angle between the chain directions at the N and C-terminus. The N-terminal chain direction is defined as the vector between residues i and $i + 4$ and the C-terminal direction as the vector between residues $j - 4$ and j.

Now, considering C_J to be the contacts of residues $\{1..i\}$ and $\{j..L\}$, the ΔG_J is set to the maximum (most positive value) of all possible i and j that fulfill the topological requirement:

$$\Delta G_J = \max_{i,j \in \{1..10\}} \left\{ \sum_{c}^{C_J} \epsilon_c + T(i + L - j)\ \delta s \right\} \tag{C.11}$$

On the other hand, considering C_C to be the contacts of the residues centered at the cut position forming the hinge loop between domains, the ΔG_C is set to the maximum (most positive value) of all possible hinge loop lengths h between the minimum L_h and maximum (set to 8 by default).

$$\Delta G_C = \min_{h \in \{L_h..8\}} \left\{ \sum_{c}^{C_C} \epsilon_c + T \cdot h \cdot \delta s \right\} \tag{C.12}$$

Model parameters

- Entropy gain of unfolding: $\delta s = 0.0054\ \mathrm{kcal/mol \cdot K \cdot residue}$
- Temperature: $T = 350\ K$
- Minimum length of the hinge loop: $L_h = 3$ residues
- Length of the inter-domain linker: $L_l = 0$ residues
- Average length contribution of a residue: $r_0 = 3.5$ Å

Appendix D

Open software and source code

This appendix contains a summary of available source code and software repositories created and used in this thesis. These resources are provided under open licenses to enable other researchers to freely reuse and modify them.

D.1 Modified T-REKS tool

https://github.com/lafita/treks-hpc

This repository contains a modified version of the T-REKS tool (Jorda and Kajava, 2009) used in Chapters 2 and 3 to detect tandem sequence repeats in proteins. It includes bug-fixes, additional user options and improved performance for large scale analyses in HPC clusters. A full list of the changes introduced can be found in the GitHub repository.

D.2 TADOSS

https://github.com/lafita/tadoss

This repository contains the source code of the TADOSS method (Lafita *et al.*, 2018), presented in Chapter 5, and can be used to calculate coarse-grained (Go-like) models from the three-dimensional structure of protein domains and to estimate the stability of their tandem domain swap conformations. Instructions to install and run the software, and results for an example input domain, can be found in the GitHub repository.

D.3 Protein EDM modelling

https://github.com/lafita/protein-edm-demo

This repository contains code written in `R` to demonstrate the protein modelling approach based on Euclidean Distance Matrices (EDMs) covered in Chapter 6 and further described in Lafita and Bateman (2020). It contains three scripts to model circular permutations, domain swaps and domain atrophy on the structure of an example SH3 domain.

D.4 Sequence composition and bias

https://github.com/bateman-research/sequence-bias

This repository contains scripts to analyse the sequence composition and bias of protein and nucleotide (DNA) sequences. It contains scripts to calculate amino acid and nucleotide composition and bias (Shannon and other entropy metrics) for a subset of sequences (for example Figures 3.5 and B.1), and to plot composition and bias profiles along a protein or DNA sequence (for example Figures 3.6 and 3.7). It also includes a script to generate dot-plots for sequence comparison and repeat detection, such as in Figure 2.1.

D.5 Sequence similarity networks and clustering

https://github.com/bateman-research/clustering-ssn

This repository contains scripts to construct Sequence Similarity Networks (SSNs) of proteins, such as the SSN of Rib domains shown in Figure 4.5, and to cluster them by sequence similarity, such as for the Periscope stalk domain repeats (Figure 3.4). It also includes a script to plot a sequence similarity matrix of protein domains in a Pfam family, shown for SHIRT domains in Figure 3.12.

Abbreviations

AFM	Atomic Force Microscopy
BLAST	Basic Local Alignment Search Tool
CASP	Critical Assessment of protein Structure Prediction techniques
CP	Circular Permutation
DNA	Deoxyribonucleic acid
DPF	Dissimilarity Parameterization Formulation
DSD	Domain Swap Dimer
EBI	European Bioinformatics Institute
ECOD	Evolutionary Classification of Protein Domains database
EDM	Euclidean Distance Matrix
EDMC	Euclidean Distance Matrix Completion
EM	Electron Microscopy
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
ESRF	European Synchrotron Radiation Facility
FRET	Förster Resonance Energy Transfer
GBS	Group B Streptoccocus
HITRD	High Identity Tandem Domain Repeat
HMM	Hidden Markov Model
HMMER	Hidden Markov Model homology search engine
MD	Molecular Dynamics
MDS	Multi-Dimensional Scaling
MBG	Mirror-Beta Grasp
NCTC	National Collection of Type Cultures
NIH	National Institutes of Health
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis

PCR Polymerase Chain Reaction
PDB Protein Data Bank
RMSD Root Mean Square Deviation
SasG *Staphylococcus aureus* surface protein G
SHIRT Streptococcal High Identity Repeats in Tandem
TADOSS TAndem DOmain Swap Stability predictor
TDS Tandem Domain Swap
T-REKS Tandem REpeat detection with a K-meanS based algorithm
TRF Tandem Repeat Finder

Bibliography

Abraham, Mark James, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindah (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2, pp. 19–25. DOI: 10.1016/j.softx.2015.06.001.

Alford, Rebecca F., Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, *et al.* (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* 13.6, pp. 3031–3048. DOI: 10.1021/acs.jctc.7b00125.

Alipanahi, Babak, Nathan Krislock, Ali Ghodsi, Henry Wolkowicz, Logan Donaldson, and Ming Li (2013). Determining protein structures from NOESY distance constraints by semidefinite programming. *Journal of Computational Biology* 20.4, pp. 296–310. DOI: 10.1089/cmb.2012.0089.

Almeida, Alexandre, Isabelle Rosinski-Chupin, Céline Plainvert, Pierre-Emmanuel Douarre, Maria J. Borrego, Claire Poyart, and Philippe Glaser (2017). Parallel Evolution of Group B Streptococcus Hypervirulent Clonal Complex 17 Unveils New Pathoadaptive Mutations. *mSystems* 2.5, pp. 00074–17. DOI: 10.1128/mSystems.00074-17.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25.17, pp. 3389–3402. DOI: 10.1093/nar/25.17.3389.

Andrade, M A, C Perez-Iratxeta, and C P Ponting (2001). Protein repeats: structures, functions, and evolution. *Journal of Structural Biology* 134.2-3, pp. 117–131. DOI: 10.1006/jsbi.2001.4392.

Apic, Gordana, Julian Gough, and Sarah A. Teichmann (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology* 310.2, pp. 311–325. DOI: 10.1006/jmbi.2001.4776.

Back, Catherine R., Victoria A. Higman, Kristian Le Vay, Viren V. Patel, Alice E. Parnell, Daniel Frankel, Howard F. Jenkinson, Steven G. Burston, Matthew P. Crump,

Angela H. Nobbs, and Paul R. Race (2020). The streptococcal multidomain fibrillar adhesin CshA has an elongated polymeric architecture. *Journal of Biological Chemistry* 295.19, pp. 6689–6699. DOI: 10.1074/jbc.RA119.011719.

Back, Catherine R., Maryta N. Sztukowska, Marisa Till, Richard J. Lamont, Howard F. Jenkinson, Angela H. Nobbs, and Paul R. Race (2017). The Streptococcus gordonii adhesin CshA protein binds host fibronectin via a catch-clamp mechanism. *Journal of Biological Chemistry* 292.5, pp. 1538–1549. DOI: 10.1074/jbc.M116.760975.

Baiesi, Marco, Enzo Orlandini, Antonio Trovato, and Flavio Seno (2016). Linking in domain-swapped protein dimers. *Scientific Reports* 6.1, p. 33872. DOI: 10.1038/srep33872.

Bashton, Matthew and Cyrus Chothia (2007). The Generation of New Protein Functions by the Combination of Domains. *Structure* 15.1, pp. 85–99. DOI: 10.1016/j.str.2006.11.009.

Bateman, Alex (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research* 47.D1, pp. D506–D515. DOI: 10.1093/nar/gky1049.

Batey, Sarah, Adrian A. Nickson, and Jane Clarke (2008). Studying the folding of multidomain proteins. *HFSP Journal* 2.6, pp. 365–377. DOI: 10.2976/1.2991513.

Bennett, Melanie J., Michael R. Sawaya, and David Eisenberg (2006). Deposition Diseases and 3D Domain Swapping. *Structure* 14.5, pp. 811–824. DOI: 10.1016/j.str.2006.03.011.

Bensing, Barbara A., Lioudmila V. Loukachevitch, Kathryn M. Mcculloch, Hai Yu, Kendra R. Vann, Zdzislaw Wawrzak, Spencer Anderson, Xi Chen, Paul M. Sullam, and T. M. Iverson (2016). Structural basis for sialoglycan binding by the streptococcus sanguinis SrpA adhesin. *Journal of Biological Chemistry* 291.14, pp. 7230–7240. DOI: 10.1074/jbc.M115.701425.

Benson, Gary (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research* 27.2, pp. 573–580. DOI: 10.1093/nar/27.2.573.

Berman, H M, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne (2000). The Protein Data Bank. *Nucleic Acids Research* 28.1, pp. 235–242. DOI: 10.1093/nar/28.1.235.

Björklund, Åsa K., Diana Ekman, and Arne Elofsson (2006). Expansion of protein domain repeats. *PLoS Computational Biology* 2.8, pp. 0959–0970. DOI: 10.1371/journal.pcbi.0020114.

Björklund, Åsa K., Diana Ekman, Sara Light, Johannes Frey-Skött, and Arne Elofsson (2005). Domain rearrangements in protein evolution. *Journal of Molecular Biology* 353.4, pp. 911–923. DOI: 10.1016/j.jmb.2005.08.067.

Bliven, Spencer and Andreas Prli (2012). Circular Permutation in Proteins. *PLoS Computational Biology* 8.3, e1002445. DOI: 10.1371/journal.pcbi.1002445.

Borgia, Alessandro, Katherine R. Kemplen, Madeleine B. Borgia, Andrea Soranno, Sarah Shammas, Bengt Wunderlich, Daniel Nettels, Robert B. Best, Jane Clarke, and Benjamin Schuler (2015). Transient misfolding dominates multidomain protein folding. *Nature Communications* 6, p. 8861. DOI: 10.1038/ncomms9861.

Borgia, Madeleine B., Alessandro Borgia, Robert B. Best, Annette Steward, Daniel Nettels, Bengt Wunderlich, Benjamin Schuler, and Jane Clarke (2011). Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* 474.7353, pp. 662–665. DOI: 10.1038/nature10099.

Borlee, Bradley R., Aaron D. Goldman, Keiji Murakami, Ram Samudrala, Daniel J. Wozniak, and Matthew R. Parsek (2010). Pseudomonas aeruginosa uses a cyclic-di-GMP-regulated adhesin to reinforce the biofilm extracellular matrix. *Molecular Microbiology* 75.4, pp. 827–842. DOI: 10.1111/j.1365-2958.2009.06991.x.

Brini, Emiliano, Carlos Simmerling, and Ken Dill (2020). Protein storytelling through physics. *Science* 370.6520. DOI: 10.1126/science.aaz3041.

Brünger, Axel T., Paul D. Adams, G. Marius Clore, Warren L. Delano, Piet Gros, Ralf W. Grosse-kunstleve, Jian Sheng Jiang, John Kuszewski, Michael Nilges, Navraj S. Pannu, Randy J. Read, Luke M. Rice, Thomas Simonson, and Gregory L. Warren (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography* 54.5, pp. 905–921. DOI: 10.1107/S0907444998003254.

Bryngelson, Joseph D D and Peter G G Wolynes (1987). Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences* 84.November, pp. 7524–7528. DOI: 10.1073/pnas.84.21.7524.

Buljan, Marija and Alex Bateman (2009). The evolution of protein domain families. *Biochemical Society Transactions* 37.4, pp. 751–755. DOI: 10.1042/BST0370751.

Bumbaca, Daniela, James E. LittleJohn, Hannah Nayakanti, Daniel J. Rigden, Michael Y. Galperin, and Mark J. Jedrzejas (2005). Sequence Analysis and Characterization of a Novel Fibronectin-Binding Repeat Domain from the Surface of Streptococcus pneumoniae. *OMICS: A Journal of Integrative Biology* 8.4, pp. 341–356. DOI: 10.1089/omi.2004.8.341.

Cheng, Hua, R. Dustin Schaeffer, Yuxing Liao, Lisa N. Kinch, Jimin Pei, Shuoyong Shi, Bong Hyun Kim, and Nick V. Grishin (2014). ECOD: An Evolutionary Classification of Protein Domains. *PLoS Computational Biology* 10.12. DOI: 10.1371/journal.pcbi.1003926.

Chothia, Cyrus, Julian Gough, Christine Vogel, and Sarah A. Teichmann (2003). Evolution of the protein repertoire. *Science* 300.5626, pp. 1701–1703. DOI: 10.1126/science.1085371.

Chothia, Cyrus and Arthur M. Lesk (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal* 5.4, pp. 823–826. DOI: 10.1002/j.1460-2075.1986.tb04288.x.

Chuang, Ya Chu, I. Chen Hu, Ping Chiang Lyu, and Shang Te Danny Hsu (2019). Untying a Protein Knot by Circular Permutation. *Journal of Molecular Biology* 431.4, pp. 857–863. DOI: 10.1016/j.jmb.2019.01.005.

Cock, Peter J.A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J.L. De Hoon (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25.11, pp. 1422–1423. DOI: 10.1093/bioinformatics/btp163.

Corrigan, Rebecca M., David Rigby, Pauline Handley, and Timothy J. Foster (2007). The role of Staphylococcus aureus surface protein SasG in adherence and biofilm formation. *Microbiology* 153.8, pp. 2435–2446. DOI: 10.1099/mic.0.2007/006676-0.

Deber, Charles M., Christopher J. Brandl, Raisa B. Deber, Lynn C. Hsu, and Xenia K. Young (1986). Amino acid composition of the membrane and aqueous domains of integral membrane proteins. *Archives of Biochemistry and Biophysics* 251.1, pp. 68–76. DOI: 10.1016/0003-9861(86)90052-4.

Deshaies, Raymond J. (2020). Multispecific drugs herald a new era of biopharmaceutical innovation. *Nature* 580.7803, pp. 329–338. DOI: 10.1038/s41586-020-2168-1.

Desvaux, Mickaël, Emilie Dumas, Ingrid Chafsey, and Michel Hébraud (2006). Protein cell surface display in Gram-positive bacteria: From single protein to macromolecular protein structure. *FEMS Microbiology Letters* 256.1, pp. 1–15. DOI: 10.1111/j.1574-6968.2006.00122.x.

Ding, Feng, Kirk C. Prutzman, Sharon L. Campbell, and Nikolay V. Dokholyan (2006). Topological determinants of protein domain swapping. *Structure* 14.1, pp. 5–14. DOI: 10.1016/j.str.2005.09.008.

Dobson, Christopher M, Tuomas P J Knowles, and Michele Vendruscolo (2019). The Amyloid Phenomenon and Its Significance in Biology and Medicine. *Cold Spring Harbor perspectives in biology*, a033878. DOI: 10.1101/cshperspect.a033878.

Doig, Andrew J. and Michael J.E. Sternberg (1995). Side-chain conformational entropy in protein folding. *Protein Science* 4.11, pp. 2247–2251. DOI: 10.1002/pro.5560041101.

Dokmanic, Ivan, Reza Parhizkar, Juri Ranieri, and Martin Vetterli (2015). Euclidean Distance Matrices: Essential theory, algorithms, and applications. *IEEE Signal Processing Magazine* 32.6, pp. 12–30. DOI: 10.1109/MSP.2015.2398954.

Dosztányi, Zsuzsanna, Veronika Csizmók, Péter Tompa, and István Simon (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology* 347.4, pp. 827–839. DOI: 10.1016/j.jmb.2005.01.071.

Drusvyatskiy, D., N. Krislock, Y. L. Voronin, and H. Wolkowicz (2017). Noisy euclidean distance realization: Robust facial reduction and the pareto frontier. *SIAM Journal on Optimization* 27.4, pp. 2301–2331. DOI: 10.1137/15M103710X.

Ebbes, Maria, Willem M. Bleymüller, Mihaela Cernescu, Rolf Nölker, Bernd Brutschy, and Hartmut H. Niemann (2011). Fold and function of the InlB B-repeat. *Journal of Biological Chemistry* 286.17, pp. 15496–15506. DOI: 10.1074/jbc.M110.189951.

Eddy, Sean R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology* 7.10, e1002195. DOI: 10.1371/journal.pcbi.1002195.

Elliott, D., E. Harrison, Pauline S. Handley, S. K. Ford, E. Jaffray, N. Mordan, and R. McNab (2003). Prevalence of Csh-like fibrillar surface proteins among mitis group oral streptococci. *Oral Microbiology and Immunology* 18.2, pp. 114–120. DOI: 10.1034/j.1399-302X.2003.00052.x.

Etzold, Sabrina, Donald A. Mackenzie, Faye Jeffers, John Walshaw, Stefan Roos, Andrew M. Hemmings, and Nathalie Juge (2014). Structural and molecular insights into novel surface-exposed mucus adhesins from lactobacillus reuteri human strains. *Molecular Microbiology* 92.3, pp. 543–556. DOI: 10.1111/mmi.12574.

Federhen, Scott (2012). The NCBI Taxonomy database. *Nucleic Acids Research* 40.D1, pp. 136–143. DOI: 10.1093/nar/gkr1178.

Fischetti, Vincent A. (2019). Surface Proteins on Gram-Positive Bacteria. *Microbiology Spectrum* 7.4, pp. 1–15. DOI: 10.1128/microbiolspec.gpp3-0012-2018.

Foster, Timothy J., Joan A. Geoghegan, Vannakambadi K. Ganesh, and Magnus Höök (2014). Adhesion, invasion and evasion: The many functions of the surface proteins of Staphylococcus aureus. *Nature Reviews Microbiology* 12.1, pp. 49–62. DOI: 10.1038/nrmicro3161.

Garcia-Manyes, Sergi, David Giganti, Carmen L Badilla, Ainhoa Lezamiz, Judit Perales-Calvo, Amy E.M. Beedle, and Julio M Fernández (2016). Single-molecule force spectroscopy predicts a misfolded, domain-swapped conformation in human γD-crystallin protein. *Journal of Biological Chemistry* 291.8, pp. 4226–4235. DOI: 10.1074/jbc.M115.673871.

Gardy, Jennifer L., Cory Spencer, Ke Wang, Martin Ester, Gábor E. Tusnády, István Simon, Sujun Hua, Katalin deFays, Christophe Lambert, Kenta Nakai, and Fiona S.L. Brinkman (2003). PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Research* 31.13, pp. 3613–3617. DOI: 10.1093/nar/gkg602.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, Lorna J. Richardson, Gustavo A. Salazar, Alfredo Smart, Erik L.L. Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C.E. Tosatto, *et al.* (2019). The Pfam protein families database in 2019. *Nucleic Acids Research* 47.D1, pp. D427–D432. DOI: 10.1093/nar/gky995.

Grant, Barry J., Ana P.C. Rodrigues, Karim M. ElSawy, J. Andrew McCammon, and Leo S.D. Caves (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics* 22.21, pp. 2695–2696. DOI: 10.1093/bioinformatics/btl461.

Gravekamp, Claudia, Debra S. Horensky, James L. Michel, and Lawrence C. Madoff (1996). Variation in repeat number within the alpha C protein of group B streptococci alters antigenicity and protective epitopes. *Infection and Immunity* 64.9, pp. 3576–3583. DOI: 10.1128/iai.64.9.3576-3583.1996.

Gräwert, Tobias W. and Dmitri I. Svergun (2020). Structural Modeling Using Solution Small-Angle X-ray Scattering (SAXS). *Journal of Molecular Biology* 432.9, pp. 3078–3092. DOI: 10.1016/j.jmb.2020.01.030.

Greener, Joe G, Nikita Desai, Shaun M Kandathil, and David T Jones (2020). Near-complete protein structural modelling of the minimal genome. *arXiv*, pp. 1–11.

Gronenborn, Angela M. (2009). Protein acrobatics in pairs - dimerization via domain swapping. *Current Opinion in Structural Biology* 19.1, pp. 39–49. DOI: 10.1016/j.sbi.2008.12.002.

Gruszka, Dominika T., Carolina A. T. F. Mendonça, Emanuele Paci, Fiona Whelan, Judith Hawkhead, Jennifer R. Potts, and Jane Clarke (2016). Disorder drives cooperative folding in a multidomain protein. *Proceedings of the National Academy of Sciences* 113.42, pp. 11841–11846. DOI: 10.1073/pnas.1608762113.

Gruszka, Dominika T., Justyna A. Wojdyla, Richard J. Bingham, Johan P. Turkenburg, Iain W. Manfield, Annette Steward, Andrew P. Leech, Joan A. Geoghegan, Timothy J. Foster, Jane Clarke, and Jennifer R. Potts (2012). Staphylococcal biofilm-forming protein has a contiguous rod-like structure. *Proceedings of the National Academy of Sciences of the United States of America* 109.17, E1011–E1018. DOI: 10.1073/pnas.1119456109.

Ha, Jeung Hoi, Joshua M. Karchin, Nancy Walker-Kopp, Carlos A. Castañeda, and Stewart N. Loh (2015). Engineered domain swapping as an on/off switch for protein function. *Chemistry and Biology* 22.10, pp. 1384–1393. DOI: 10.1016/j.chembiol.2015.09.007.

Han, Jung Hoon, Sarah Batey, Adrian A. Nickson, Sarah A. Teichmann, and Jane Clarke (2007). The folding and evolution of multidomain proteins. *Nature Reviews Molecular Cell Biology* 8.4, pp. 319–330. DOI: 10.1038/nrm2144.

Harrison, Paul M. and Mark Gerstein (2003). A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes. *Genome biology* 4.6. DOI: 10.1186/gb-2003-4-6-r40.

Heger, Andreas and Liisa Holm (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins: Structure, Function and Genetics* 41.2, pp. 224–237. DOI: 10.1002/1097-0134(20001101)41:2<224::AID-PROT70>3.0.CO;2-Z.

Hu, Zengjian, Donnell Bowen, William M. Southerland, Antonio Del Sol, Yongping Pan, Ruth Nussinov, and Buyong Ma (2007). Ligand binding and circular permutation modify residue interaction network in DHFR. *PLoS Computational Biology* 3.6, pp. 1097–1107. DOI: 10.1371/journal.pcbi.0030117.

Hubbard, Tim J.P., Bart Ailey, Steven E. Brenner, Alexey G. Murzin, and Cyrus Chothia (1999). SCOP: A structural classification of proteins database. DOI: 10.1093/nar/27.1.254.

Iwakura, Masahiro, Tsutomu Nakamura, Chiori Yamane, and Kosuke Maki (2000). Systematic circular permutation of an entire protein reveals essential folding elements. *Nature Structural Biology* 7.7, pp. 580–585. DOI: 10.1038/76811.

Jacob, François (1977). Evolution and tinkering. *Science* 196.4295, pp. 1161–1166. DOI: 10.1126/science.860134.

Jarnot, Patryk, Joanna Ziemska-Legiecka, Laszlo Dobson, Matthew Merski, Pablo Mier, Miguel A. Andrade-Navarro, John M. Hancock, Zsuzsanna Dosztányi, Lisanna Paladin, Marco Necci, Damiano Piovesan, Silvio C.E. Tosatto, Vasilis J. Promponas, Marcin Grynberg, and Aleksandra Gruca (2020). PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. *Nucleic acids research* 48.W1, W77–W84. DOI: 10.1093/nar/gkaa339.

Jensch, Inga, Gustavo Gámez, Michael Rothe, Sandra Ebert, Marcus Fulde, Daniela Somplatzki, Simone Bergmann, Lothar Petruschka, Manfred Rohde, Roland Nau, and Sven Hammerschmidt (2010). PavB is a surface-exposed adhesin of Streptococcus pneumoniae contributing to nasopharyngeal colonization and airways infections. *Molecular Microbiology* 77.1, pp. 22–43. DOI: 10.1111/j.1365-2958.2010.07189.x.

Jonic, Slavica and Catherine Vénien-Bryan (2009). Protein structure determination by electron cryo-microscopy. DOI: 10.1016/j.coph.2009.04.006.

Jorda, Julien and Andrey V. Kajava (2009). T-REKS: Identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* 25.20, pp. 2632–2638. DOI: 10.1093/bioinformatics/btp482.

Karanicolas, John and Charles L. Brooks (2002). The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Science* 11.10, pp. 2351–2361. DOI: 10.1110/ps.0205402.

Kishan, K.V. Radha, Marcia E. Newcomer, Thomas H. Rhodes, and Stephen D. Guilliot (2001). Effect of pH and salt bridges on structural assembly: Molecular structures of the monomer and intertwined dimer of the Eps8 SH3 domain. *Protein Science* 10.5, pp. 1046–1055. DOI: 10.1110/ps.50401.

Kobe, Bostjan and Johann Deisenhofer (1994). The leucine-rich repeat: a versatile binding motif. *Trend in Biochemical Science* 19.October, pp. 415–421.

Lachenauer, C S, R Creti, J L Michel, and L C Madoff (2000). Mosaicism in the alpha-like protein genes of group B streptococci. *Proc Natl Acad Sci U S A* 97.17, pp. 9630–5. DOI: 10.1073/pnas.97.17.9630.

Lafita, Aleix and Alex Bateman (2020). Modelling structural rearrangements in proteins using Euclidean distance matrices. *F1000Research* 9.728, pp. 1–7. DOI: 10.12688/f1000research.25235.1.

Lafita, Aleix, Pengfei Tian, Robert B. Best, and Alex Bateman (2018). TADOSS: computational estimation of tandem domain swap stability. *Bioinformatics* 35.14, pp. 1–2. DOI: 10.1093/bioinformatics/bty974.

— (2019). Tandem domain swapping: determinants of multidomain protein misfolding. *Current Opinion in Structural Biology* 58, pp. 97–104. DOI: 10.1016/j.sbi.2019.05.012.

Leman, Julia Koehler, Brian D. Weitzner, Steven M. Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F. Alford, Melanie Aprahamian, David Baker, Kyle A. Barlow, Patrick Barth, Benjamin Basanta, Brian J. Bender, Kristin Blacklock, Jaume Bonet, Scott E. Boyken, *et al.* (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods* 17.7, pp. 665–680. DOI: 10.1038/s41592-020-0848-2.

Letunic, Ivica and Peer Bork (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research* 47.W1, W256–W259. DOI: 10.1093/nar/gkz239.

Levitt, Michael (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America* 106.27, pp. 11079–11084. DOI: 10.1073/pnas.0905029106.

Li, Jing, Jing Wen Li, Zhixing Feng, Juanjuan Wang, Haoran An, Yanni Liu, Yang Wang, Kailing Wang, Xuegong Zhang, Zhun Miao, Wenbo Liang, Robert Sebra, Guilin Wang, Wen Ching Wang, and Jing Ren Zhang (2016). Epigenetic Switch Driven by DNA Inversions Dictates Phase Variation in Streptococcus pneumoniae. *PLoS Pathogens* 12.7, pp. 1–36. DOI: 10.1371/journal.ppat.1005762.

Li, Pilong, Sudeep Banjade, Hui Chun Cheng, Soyeon Kim, Baoyu Chen, Liang Guo, Marc Llaguno, Javoris V. Hollingsworth, David S. King, Salman F. Banani, Paul S. Russo, Qiu Xing Jiang, B. Tracy Nixon, and Michael K. Rosen (2012). Phase tran-

sitions in the assembly of multivalent signalling proteins. *Nature* 483.7389, pp. 336–340. DOI: 10.1038/nature10879.

Lim, Kian Guan, Chee Keong Kwoh, Li Yang Hsu, and Adrianto Wirawan (2013). Review of tandem repeat search tools: A systematic approach to evaluating algorithmic performance. *Briefings in Bioinformatics* 14.1, pp. 67–81. DOI: 10.1093/bib/bbs023.

Liu, Lin, In Ja L. Byeon, Ivet Bahar, and Angela M. Gronenborn (2012). Domain swapping proceeds via complete unfolding: A 19F- and 1H-NMR study of the cyanovirin-N protein. *Journal of the American Chemical Society* 134.9, pp. 4229–4235. DOI: 10.1021/ja210118w.

Lo, Wei Cheng, Tian Dai, Yen Yi Liu, Li Fen Wang, Jenn Kang Hwang, and Ping Chiang Lyu (2012a). Deciphering the preference and predicting the viability of circular permutations in proteins. *PLoS ONE* 7.2, e31791. DOI: 10.1371/journal.pone.0031791.

Lo, Wei Cheng, Chi Ching Che Yu Lee, Chi Ching Che Yu Lee, and Ping Chiang Lyu (2009). CPDB: A database of circular permutation in proteins. *Nucleic Acids Research* 37, pp. D328–D332. DOI: 10.1093/nar/gkn679.

Lo, Wei Cheng, Li Fen Wang, Yen Yi Liu, Tian Dai, Jenn Kang Hwang, and Ping Chiang Lyu (2012b). CPred: A web server for predicting viable circular permutations in proteins. *Nucleic Acids Research* 40.W1, W232–W237. DOI: 10.1093/nar/gks529.

Lupas, A N, C P Ponting, and R B Russell (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of structural biology* 134.2-3, pp. 191–203. DOI: 10.1006/jsbi.2001.4393.

Malevanets, Anatoly, Fernanda L. Sirota, and Shoshana J. Wodak (2008). Mechanism and Energy Landscape of Domain Swapping in the B1 Domain of Protein G. *Journal of Molecular Biology* 382.1, pp. 223–235. DOI: 10.1016/j.jmb.2008.06.025.

Markowetz, Florian (2017). All biology is computational biology. *PLoS Biology* 15.3, pp. 4–7. DOI: 10.1371/journal.pbio.2002050.

Marks, Debora S., Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6.12. DOI: 10.1371/journal.pone.0028766.

Mascarenhas, Nahren Manuel and Shachi Gosavi (2017). Understanding protein domain-swapping using structure-based models of protein folding. *Progress in Biophysics and Molecular Biology* 128, pp. 113–120. DOI: 10.1016/j.pbiomolbio.2016.09.013.

McNab, Roderick, Helen Forbes, Pauline S. Handley, Diane M. Loach, Gerald W. Tannock, and Howard F. Jenkinson (1999). Cell wall-anchored csha polypeptide (259 kilodaltons) in streptococcus gordonii forms surface fibrils that confer hydrophobic

and adhesive properties. *Journal of Bacteriology* 181.10, pp. 3087–3095. DOI: 10.1128/jb.181.10.3087-3095.1999.

Melia, Charlotte E, Jani R Bolla, Stefan Katharios-lanwermeyer, Daniel B Mihaylov, Patrick C Hoffmann, Jiandong Huo, Michael R Wozny, Louis M Elfari, Raymond J Owens, Carol V Robinson, and Tanmay A M Bharat (2021). Architecture of cell-cell junctions in situ reveals a mechanism for bacterial biofilm inhibition. *bioRxiv*, pp. 1–28. DOI: 10.1101/2021.02.08.430230.

Mier, Pablo, Lisanna Paladin, Stella Tamana, Sophia Petrosian, Borbála Hajdu-Soltész, Annika Urbanek, Aleksandra Gruca, Dariusz Plewczynski, Marcin Grynberg, Pau Bernadó, Zoltán Gáspári, Christos A. Ouzounis, Vasilis J. Promponas, Andrey V. Kajava, John M. Hancock, *et al.* (2020). Disentangling the complexity of low complexity proteins. *Briefings in Bioinformatics* 21.2, pp. 458–472. DOI: 10.1093/bib/bbz007.

Miller, J., A. D. McLachlan, and A. Klug (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO journal* 4.6, pp. 1609–1614. DOI: 10.1002/j.1460-2075.1985.tb03825.x.

Mitchell, Alex L, Maxim Scheremetjew, Hubert Denise, Simon Potter, Aleksandra Tarkowska, Matloob Qureshi, Gustavo A Salazar, Sebastien Pesseat, Miguel A Boland, Fiona M.I. Hunter, Petra Ten Hoopen, Blaise Alako, Clara Amid, Darren J Wilkinson, Thomas P Curtis, *et al.* (2018). EBI Metagenomics in 2017: Enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Research* 46.D1, pp. D726–D735. DOI: 10.1093/nar/gkx967.

Miyazawa, Sanzo and Robert L. Jernigan (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology* 256.3, pp. 623–644. DOI: 10.1006/jmbi.1996.0114.

Monzon, Vivian, Aleix Lafita, and Alex Bateman (2020). Discovery of fibrillar adhesins across bacterial species. *bioRxiv*, pp. 1–31. DOI: 10.1101/2020.12.07.414375.

Moult, John, Jan T. Pedersen, Richard Judson, and Krzysztof Fidelis (1995). A largescale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics* 23.3, pp. ii–iv. DOI: 10.1002/prot.340230303.

Moxon, Richard, Chris Bayliss, and Derek Hood (2006). Bacterial contingency loci: The role of simple sequence DNA repeats in bacterial adaptation. *Annual Review of Genetics* 40, pp. 307–333. DOI: 10.1146/annurev.genet.40.110405.090442.

Nandwani, Neha, Parag Surana, Hitendra Negi, Nahren M. Mascarenhas, Jayant B. Udgaonkar, Ranabir Das, and Shachi Gosavi (2019). A five-residue motif for the design of domain swapping in proteins. *Nature Communications* 10.1, p. 452. DOI: 10.1038/s41467-019-08295-x.

Navarre, William Wiley and Olaf Schneewind (1999). Surface Proteins of Gram-Positive Bacteria and Mechanisms of Their Targeting to the Cell Wall Envelope. DOI: 1092-2172/99/{\$}04.00+0.

Nylander, Åsa, Gunnel Svensäter, Dilani B. Senadheera, Dennis G. Cvitkovitch, Julia R. Davies, and Karina Persson (2013). Structural and Functional Analysis of the N-terminal Domain of the Streptococcus gordonii Adhesin Sgo0707. *PLoS ONE* 8.5. DOI: 10.1371/journal.pone.0063768.

Oberhauser, Andres F., Piotr E. Marszalek, Mariano Carrion-Vazquez, and Julio M. Fernandez (1999). Single protein misfolding events captured by atomic force microscopy. *Nature Structural Biology* 6.11, pp. 1025–1028. DOI: 10.1038/14907.

Paladin, Lisanna, Layla Hirsh, Damiano Piovesan, Miguel A. Andrade-Navarro, Andrey V. Kajava, and Silvio C.E. Tosatto (2017). RepeatsDB 2.0: Improved annotation, classification, search and visualization of repeat protein structures. *Nucleic Acids Research* 45.D1, pp. D308–D312. DOI: 10.1093/nar/gkw1136.

Pancsa, Rita, Denes Kovacs, and Peter Tompa (2019). Misprediction of structural disorder in halophiles. *Molecules* 24.3. DOI: 10.3390/molecules24030479.

Paszkiewicz, Konrad H., Michael J.E. Sternberg, and Michael Lappe (2006). Prediction of viable circular permutants using a graph theoretic approach. *Bioinformatics* 22.11, pp. 1353–1358. DOI: 10.1093/bioinformatics/btl095.

Pellegrini, Marco (2015). Tandem Repeats in Proteins: Prediction Algorithms and Biological Role. *Frontiers in Bioengineering and Biotechnology* 3. DOI: 10.3389/fbioe.2015.00143.

Perez-Riba, Albert and Laura S. Itzhaki (2017). A method for rapid high-throughput biophysical analysis of proteins. *Scientific Reports* 7.1, pp. 1–6. DOI: 10.1038/s41598-017-08664-w.

Perry, Andrew J and Bosco K Ho (2013). Inmembrane, a bioinformatic workflow for annotation of bacterial cell-surface proteomes. *Source Code for Biology and Medicine* 8.1, p. 9. DOI: 10.1186/1751-0473-8-9.

Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 25.13, pp. 1605–1612. DOI: 10.1002/jcc.20084.

Phillips, Zachary N., Greg Tram, Kate L. Seib, and John M. Atack (2019). Phase-variable bacterial loci: How bacteria gamble to maximise fitness in changing environments. *Biochemical Society Transactions* 47.4, pp. 1131–1141. DOI: 10.1042/BST20180633.

Pickett, Stephen D. and Michael J.E. Sternberg (1993). Empirical scale of side-chain conformational entropy in protein folding. DOI: 10.1006/jmbi.1993.1329.

Prakash, Ananth and Alex Bateman (2015). Domain atrophy creates rare cases of functional partial protein domains. *Genome Biology* 16.1, pp. 1–15. DOI: 10.1186/s13059-015-0655-8.

Pyburn, Tasia M., Barbara A. Bensing, Yan Q. Xiong, Bruce J. Melancon, Thomas M. Tomasiak, Nicholas J. Ward, Victoria Yankovskaya, Kevin M. Oliver, Gary Cecchini, Gary A. Sulikowski, Matthew J. Tyska, Paul M. Sullam, and T. M. Iverson (2011). A structural model for binding of the serine-rich repeat adhesin gspb to host carbohydrate receptors. *PLoS Pathogens* 7.7, pp. 6–9. DOI: 10.1371/journal.ppat.1002112.

Rahman, Adam and Wayne Oldford (2016). edmcr - Euclidean Distance Matrix Completion in R. *Journal of Statistical Software*, pp. 1–40. DOI: 10.18637/jss.v000.i00.

Reed, Christopher J., Hunter Lewis, Eric Trejo, Vern Winston, and Caryn Evilia (2013). Protein adaptations in archaeal extremophiles. *Archaea* 2013. DOI: 10.1155/2013/373275.

Roche, Fiona M., Ruth Massey, Sharon J. Peacock, Nicholas P.J. Day, Livia Visai, Pietro Speziale, Alex Lam, Mark Pallen, and Timothy J. Foster (2003). Characterization of novel LPXTG-containing proteins of Staphylococcus aureus identified from genome sequences. *Microbiology* 149.3, pp. 643–654. DOI: 10.1099/mic.0.25996-0.

Rousseau, Frederic, J. W.H. Schymkowitz, H. R. Wilkinson, and L. S. Itzhaki (2001). Three-dimensional domain swapping in p13suc1 occurs in the unfolded state and is controlled by conserved proline residues. *Proceedings of the National Academy of Sciences of the United States of America* 98.10, pp. 5596–5601. DOI: 10.1073/pnas.101542098.

Rousseau, Frederic, Joost Schymkowitz, and Laura S. Itzhaki (2012). Implications of 3D domain swapping for protein folding, misfolding and function. *Advances in Experimental Medicine and Biology* 747, pp. 137–152. DOI: 10.1007/978-1-4614-3229-6{_}9.

Schaffer, Miroslava, Stefan Pfeffer, Julia Mahamid, Stephan Kleindiek, Tim Laugks, Sahradha Albert, Benjamin D. Engel, Andreas Rummel, Andrew J. Smith, Wolfgang Baumeister, and Juergen M. Plitzko (2019). A cryo-FIB lift-out technique enables molecular-resolution cryo-ET within native Caenorhabditis elegans tissue. *Nature Methods* 16.8, pp. 757–762. DOI: 10.1038/s41592-019-0497-5.

Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin ídek, Alexander W.R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, *et al.* (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577.7792, pp. 706–710. DOI: 10.1038/s41586-019-1923-7.

Shameer, Khader, Prashant N. Shingate, S. C.P. Manjunath, M. Karthika, Ganesan Pugalenthi, and Ramanathan Sowdhamini (2011). 3DSwap: Curated knowledgebase of proteins involved in 3D domain swapping. *Database* 2011.0, bar042–bar042. DOI: 10.1093/database/bar042.

Shindyalov, I N and P E Bourne (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering Design and Selection* 11.9, pp. 739–747. DOI: 10.1093/protein/11.9.739.

Sillitoe, Ian, Tony E. Lewis, Alison Cuff, Sayoni Das, Paul Ashford, Natalie L. Dawson, Nicholas Furnham, Roman A. Laskowski, David Lee, Jonathan G. Lees, Sonja Lehtinen, Romain A. Studer, Janet Thornton, and Christine A. Orengo (2015). CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research* 43.D1, pp. D376–D381. DOI: 10.1093/nar/gku947.

Smyth, M S and J H J Martin (2000). Review x Ray crystallography. *J Clin Pathol: Mol Pathol* 53.1, pp. 8–14. DOI: 10.1136/mp.53.1.8.

Sonnhammer, Erik L.L., Sean R. Eddy, Ewan Birney, Alex Bateman, and Richard Durbin (1998). Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research* 26.1, pp. 320–322. DOI: 10.1093/nar/26.1.320.

Tian, Pengfei and Robert B. Best (2016). Structural Determinants of Misfolding in Multidomain Proteins. *PLoS Computational Biology* 12.5. DOI: 10.1371/journal.pcbi.1004933.

Tørresen, Ole K., Bastiaan Star, Pablo Mier, Miguel A. Andrade-Navarro, Alex Bateman, Patryk Jarnot, Aleksandra Gruca, Marcin Grynberg, Andrey V. Kajava, Vasilis J. Promponas, Maria Anisimova, Kjetill S. Jakobsen, and Dirk Linke (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic acids research* 47.21, pp. 10994–11006. DOI: 10.1093/nar/gkz841.

Treangen, Todd J. and Steven L. Salzberg (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics* 13.1, pp. 36–46. DOI: 10.1038/nrg3117.

Trosset, Michael W. (2000). Distance matrix completion by numerical optimization. *Computational Optimization and Applications* 17.1, pp. 11–22. DOI: 10.1023/A:1008722907820.

Uliel, S., A. Fliess, and R. Unger (2002). Naturally occurring circular permutations in proteins. *Protein Engineering Design and Selection* 14.8, pp. 533–542. DOI: 10.1093/protein/14.8.533.

Velankar, Sameer, José M. Dana, Julius Jacobsen, Glen Van Ginkel, Paul J. Gane, Jie Luo, Thomas J. Oldfield, Claire O'Donovan, Maria Jesus Martin, and Gerard J. Kleywegt (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research* 41.D1, pp. 483–489. DOI: 10.1093/nar/gks1258.

Vink, Cornelis, Gloria Rudenko, and H. Steven Seifert (2012). Microbial antigenic variation mediated by homologous DNA recombination. *FEMS Microbiology Reviews* 36.5, pp. 917–948. DOI: 10.1111/j.1574-6976.2011.00321.x.

Vogel, Christine, Matthew Bashton, Nicola D. Kerrison, Cyrus Chothia, and Sarah A. Teichmann (2004). Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology* 14.2, pp. 208–216. DOI: 10.1016/j.sbi.2004.03.011.

Wästfelt, Maria, Margaretha Stålhammar-Carlemalm, Anne Marie Delisse, Teresa Cabezon, and Gunnar Lindahl (1996). Identification of a family of Streptococcal surface proteins with extremely repetitive structure. *Journal of Biological Chemistry* 271.31, pp. 18892–18897. DOI: 10.1074/jbc.271.31.18892.

Waterhouse, Andrew M., James B. Procter, David M.A. Martin, Michèle Clamp, and Geoffrey J. Barton (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25.9, pp. 1189–1191. DOI: 10.1093/bioinformatics/btp033.

Waudby, Christopher A., Christopher M. Dobson, and John Christodoulou (2019). Nature and Regulation of Protein Folding on the Ribosome. *Trends in Biochemical Sciences* 44.11, pp. 914–926. DOI: 10.1016/j.tibs.2019.06.008.

Wheeler, Travis J. and Sean R. Eddy (2013). Nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29.19, pp. 2487–2489. DOI: 10.1093/bioinformatics/btt403.

Whelan, Fiona, Aleix Lafita, James Gilburt, Clément Dégut, Samuel C. Griffiths, Huw T. Jenkins, Alexander N. St John, Emanuele Paci, James W.B. Moir, Michael J. Plevin, Christoph G. Baumann, Alex Bateman, and Jennifer R. Potts (2020). Periscope proteins are variable length regulators of bacterial cell surface interactions. *bioRxiv*, pp. 1–28. DOI: 10.1101/2020.12.24.424174.

Whelan, Fiona, Aleix Lafita, Samuel C. Griffiths, Rachael E.M. Cooper, Jean L. Whittingham, Johan P. Turkenburg, Iain W. Manfield, Alexander N. St John, Emanuele Paci, Alex Bateman, and Jennifer R. Potts (2019). Defining the remarkable structural malleability of a bacterial surface protein Rib domain implicated in infection. *Proceedings of the National Academy of Sciences of the United States of America* 116.52, pp. 26540–26548. DOI: 10.1073/pnas.1911776116.

Word, J. Michael, Simon C. Lovell, Jane S. Richardson, and David C. Richardson (1999). Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology* 285.4, pp. 1735–1747. DOI: 10.1006/jmbi.1998.2401.

Wright, Caroline F., Sarah A. Teichmann, Jane Clarke, and Christopher M. Dobson (2005). The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* 438.7069, pp. 878–881. DOI: 10.1038/nature04195.

Wüthrich, K. (2001). The way to NMR structures of proteins. *Nature Structural Biology* 8.11, pp. 923–925. DOI: 10.1038/nsb1101-923.

Xu, Dong and Ruth Nussinov (1998). Favorable domain size in proteins. *Folding and Design* 3.1, pp. 11–17. DOI: 10.1016/S1359-0278(98)00004-2.

Yang, Fan, Carole A. Bewley, John M. Louis, Kirk R. Gustafson, Michael R. Boyd, Angela M. Gronenborn, G. Marius Clore, and Alexander Wlodawer (1999). Crystal structure of cyanovirin-N, a potent HIV-inactivating protein, shows unexpected domain swapping. *Journal of Molecular Biology* 288.3, pp. 403–412. DOI: 10.1006/jmbi.1999.2693.

Yang, Jianyi, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences of the United States of America* 117.3, pp. 1496–1503. DOI: 10.1073/pnas.1914677117.

Yang, Sichun, Samuel S. Cho, Yaakov Levy, Margaret S. Cheung, Herbert Levine, Peter G. Wolynes, and José N. Onuchic (2004). Domain swapping is a consequence of minimal frustration. *Proceedings of the National Academy of Sciences* 101.38, pp. 13786–13791. DOI: 10.1073/pnas.0403724101.

Zheng, Weihua, Nicholas P. Schafer, and Peter G. Wolynes (2013). Frustration in the energy landscapes of multidomain protein misfolding. *Proceedings of the National Academy of Sciences* 110.5, pp. 1680–1685. DOI: 10.1073/pnas.1222130110.

Zhou, Kai, Abram Aertsen, and Chris W. Michiels (2014). The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiology Reviews* 38.1, pp. 119–141. DOI: 10.1111/1574-6976.12036.