A domain based protein structural modelling platform applied in the analysis of alternative splicing

Su Datt Lam

A thesis submitted for the degree of

Doctor of Philosophy

January 2018

Institute of Structural and Molecular Biology

University College London

Declaration

I, Su Datt Lam, confirm that the work presented in this thesis is my own except where otherwise stated. Where the thesis is based on work done by myself jointly with others or where information has been derived from other sources, this has been indicated in the thesis.

Su Datt Lam

January 2018

Abstract

Functional families (FunFams) are a sub-classification of CATH protein domain superfamilies that cluster relatives likely to have very similar structures and functions. The functional purity of FunFams has been demonstrated by comparing against experimentally determined Enzyme Commission annotations and by checking whether known functional sites coincide with highly conserved residues in the multiple sequence alignments of FunFams. We hypothesised that clustering relatives into FunFams may help in protein structure modelling.

In the first work chapter, we demonstrate the structural coherence of domains in FunFams. We then explore the usage of FunFams in protein monomer modelling. The FunFam based protocol produced higher percentages of good models compared to an HHsearch (the state-of-the-art HMM based sequence search tool) based protocol for both close and remote homologs. We developed a modelling pipeline that, utilises the FunFam protocol, and is able to model up to 70% of domain sequences from human and fly genomes.

In the second work chapter, we explore the usage of FunFams in protein complex modelling. Our analysis demonstrated that domain-domain interfaces in FunFams tend to be conserved. The FunFam based complex modelling protocol produced significantly more good quality models when compared to a BLAST based protocol and slightly better than a HHsearch based protocol.

In the final work chapter, we employ the FunFam based structural modelling tool to understand the implications of alternative splicing. We focused on isoforms derived from mutually exclusively exons (MXEs) for which there is more enriched in proteomics data. MXEs which could be mapped to structure show a significant tendency to be exposed to the solvent, are likely to exhibit a significant change in their physiochemical property and to lie close to a known/predicted functional sites. Our results suggest that MXE events may have a number of important roles in cells generally.

Acknowledgements

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

It has been an amazing experience been able to pursue my studies at the University College London. I have gained a lot of scientific knowledge from my research degree and professional development. I would like to thank those that have been instrumental in helping me achieve and reach this stage.

First and for most, I would like to express my deepest gratitude to Prof. Christine Orengo, my main supervisor. Thank you for your willingness to accept me to study and work with such a great team of people in your group. All your encouragement, guidance and support has been the powerhouse to keep me motivated throughout these four years. I will definitely miss all those weekly tutorial sessions we had. Thank you for being such a good mentor, Christine.

I would like to thank the Malaysia's Ministry of Education and the National University of Malaysia for their financial support. I would like to thank my Thesis committee members, Prof. David Jones, Dr. Konstantinos Thalassinos and Dr. Joanne Santini for their invaluable insight and support during my thesis committee meetings.

Special thanks to Dr. Sayoni Das for her constant help and motivation throughout my PhD. Also to Dr. Jonathan Lees for the suggestions and ideas that had driven me to venture into alternative splicing. Appreciation to Dr. Ian Sillitote, Dr. Tony Lewis, Dr. Natalie Dawson and Dr. David Lee for curating the CATH resource and providing computational support when I needed it. Also would like to thank Dr. Paul Ashford for his insight on structural clustering and Dr. Aurelio Moya Garcia for direction and advice on developing my protein complex modelling pipeline and providing the computational scripts to predict allosteric sites. In addition, thanks to all the other members of the Orengo group especially Camilla Pang and Tolulope Adeyelu for always being kind and helpful when I needed it.

I also like to thank all my amazing friends when I am off college. I am deeply grateful to Susanna, Jing Wei, Zora for being such an amazing and warm friends. I am also in debt to my fellow housemates/friends, Filsan, Jonas, Alvin, Bruna, Peter, Wagner, Jaimie for all the wonderful times spent and being there for me all the time. My friends back in Malaysia, Yi Ling, Ai Cheng, Beng Si and Lynice, for their best wishes throughout.

Last but not the least, heartfelt thanks to my parents, my brothers, my sisters-in-law for their unwavering love and support at the most difficult times.

Su Datt

September 2017

Contents

List of Figures

List of Tables

List of Abbreviations

3D Three-dimensional

AS Alternative Splicing

BLAST Basic Local Alignment Search Tool

BLOSUM Blocks Substitution Matrices

CASP Critical Assessment of Methods of Protein Structure Prediction

CAPRI Critical Assessment of Predicted Interactions

CAFA Critical Assessment of Protein Function Annotation

CATH Class, Architecture, Topology, Homologous Superfamily

COSMIC Catalogue Of Somatic Mutations In Cancer

Cryo-EM Cryo-electron microscopy

CSA Catalytic Site Atlas

DAVID Database for Annotation, Visualization and Integrated Discovery

DDI Domain-domain interactions

DNA Deoxyribonucleic acid

DP Dynamic Programming

DSCAM Drosophila Down Syndrome Cell Adhesion Molecule gene

DOPS Diversity of Position Score

DOPE Discrete Optimized Protein Energy

EC Enzyme Commission

E-value Expectation value

FFT Fast Fourier transform

Fnat Fraction of Native residue contacts

Fnon-nat Fraction of non-native contacts

FunFam Functional family

FunMod Functional family modelling platform

GDT-TS Global Distance test

GDT-HA Global Distance test (high accuracy version)

GO Gene Ontology

HMM Hidden Markov Model

I3D Interactome3D

IBIS Inferred Biological Interactions Server

L1CAM L1 cell adhesion molecule

MDA Multi-domain Architecture

MQAP Model Quality Assessment Protocols

mRNA messenger Ribonucleic acid

MS Mass Spectrometry

MSA Multiple sequence alignment

MXE Mutually Exclusively Exon

NCBI National Center for Biotechnology Information

NMR Nuclear Magnetic Resonance Spectroscopy

PAM Dayhoff Point Accepted Mutation

PBD4 Protein-protein Docking Benchmark version 4 dataset

PBD5 Protein-protein Docking Benchmark version 5 dataset

PDB Protein Data Bank

PKM Pyruvate Kinase

PPI Protein-protein interaction

PSI-BLAST Position Specific Iterated-BLAST

PSI Protein Structure Initiative

PSSM Position-Specific Scoring Matrix

rASA relative Accessible Surface Area

RMSD Root Mean Squared Deviations

RNA-seq RNA sequencing

S90 sequence identity 90% cluster

S35 sequence identity 35% cluster

S30 sequence identity 30% cluster

SASA Solvent Accessible Surface Area

SCOP Structural Classification of Proteins

SFLD Structure Function Linkage Database

SSAP Sequential Structure Alignment Program

SVM Support Machine Vector

TM-score Template Modeling score

UniProt Universal Protein resource

UniProtKB UniProt Knowledgebase

X-ray X-ray crystallography

List of Publications

Tony Lewis, Ian Sillitoe, Natalie L Dawson, **Su Datt Lam**, Sayoni Das, Tristan Clarke, David Lee, Christine Orengo, Jonathan Lees (2018). Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Research*, 46(D1):D435-D439.

Su Datt Lam, Sayoni Das, Ian Sillitoe, and Christine Orengo (2017). An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences. *Acta Crystallographica Section D: Structural Biology*, 73(8):628-640.

Natalie L Dawson, Ian Sillitoe, Jonathan G Lees, **Su Datt Lam**, and Christine A Orengo (2017). CATH-Gene3D: Generation of the resource and its use in obtaining structural and functional annotations for protein sequences. *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*, pages 79-110.

Su Datt Lam, Natalie L Dawson, Sayoni Das, Ian Sillitoe, Paul Ashford, David Lee, Sonja Lehtinen, Christine A Orengo, and Jonathan G Lees (2016). Gene3D: expanding the utility of domain assignments. *Nucleic Acids Research*, 44(D1):D404-D409.

Chapter 1

Introduction

Proteins are essential macromolecules found throughout biological systems. They are involved in virtually every biological process. For example, antibodies are proteins produced by our body after exposure to foreign particles (e.g. viruses and bacteria) that help neutralize the toxins and protect the body. Immunoglobulin A is an antibody found in mucosal areas, such as the gut and respiratory tract and play a crucial role in the immune function.

Enzymes are proteins that act as catalysts for most biochemical interactions in the cell. They are also involved in facilitating the process of formation of new molecules. Alcohol dehydrogenase is involved in breaking down alcohols producing different useful ketones and aldehydes. For example, in humans, alcohol dehydrogenase is found in liver and catalyses the oxidation of ethanol to acetaldehyde.

Messenger proteins, such as hormones, transmit signals to coordinate biological processes (i.e. growth, metabolism) between different cells, tissues and organs. Structural proteins, such as actin, keratin and collagen provide structure and support for cells. Last but not least, there are storage and transport proteins that bind and carry molecules throughout the body.

In order to understand biological systems, we need to know how these proteins fold and interact with other molecules (i.e. proteins, ligands). This requires the use of protein structure data. Once a protein structure is determined, it is usually deposited in the Protein Data Bank (PDB) (Berman et al., 2000). In Aug 2017, there were more than 130,000 entries in the PDB. However, solving the structure of a protein is complicated and time-consuming, therefore the number of solved protein structures is much less than the number of proteins sequenced so far, which is $\geq$89 million (UniProt, 2017). This gap is known as the "Protein Structure Gap". Various computational protein structure modelling approaches have been developed to bridge the gap.

In this thesis, work is presented which describes the development of protein monomer and protein complex structural modelling protocols, aided by template selection from functionally coherent protein families in the CATH-Gene3D classification of domain families. We then applied the structural modelling platform to understand the structural and functional implications of alternative splicing. This chapter describes general bioinformatics concepts, methods and resources that are relevant to the work presented in this thesis, followed by an outline of the following chapters.

1.1 Experimental techniques to determine the protein structure

The structures of proteins can be determined using experimental techniques such as X-ray crystallography (X-ray), nuclear magnetic resonance spectroscopy (NMR), and cryo-electron microscopy (Cryo-EM). Most of the structures that are available in the PDB database were determined using X-ray crystallography. Proteins need to be purified and crystallized before being subjected to bombardment by an intense beam of X-rays. The proteins diffract the X-ray beam into a pattern of spots, which are then analysed to determine the distribution of electrons, which is subsequently analysed to determine the location of every atom in the protein. X-ray determination provides great detail on the location of all atoms in the protein together with other ligands that were incorporated into the crystal. Nowadays, the imaging process is straightforward and crystallographers spend most of their time in the complicated process of obtaining crystals. Success in this depends on satisfying different factors such as protein purity, pH, the concentration of protein, temperature, precipitants and additives.

The accuracy of the 3D structure is dependent on the quality of crystals, which should be well ordered. An important measure of the accuracy is the X-ray resolution, which measures the amount of detail that can be seen: the higher the resolution, the higher quality of the protein structure. Proteins with an X-ray resolution below 1Å are very well determined and for most analyses, a resolution better than 4Å is desirable.

NMR spectroscopy is also used to determine protein structures. One major advantage of NMR is its ability to determine the structure of the proteins in solution, as opposed to X-ray crystallography that examines proteins in the crystal form. After the proteins are purified, they are placed in a strong magnetic field that probes with radio waves. Based on the observed resonance, the atoms that are close to each other (distance restraints) and how they are bonded (angle restraints and orientation restraints) can be deduced. An expert can then build the models based on the list of constraints. NMR can generally be applied to small proteins (bigger than 30 kDa). Large proteins tend to have too many peaks in the NMR spectra that are overlapping which complicates the structure determination process. The introduction of techniques such as isotope labelling and multidimensional experiments may alleviate these problems to some extent (Verardi et al., 2012).

NMR analyses usually include a whole ensemble of structures that are built based on the observed list of experimental restraints. Regions with strong restraints will be very similar among the structures in the ensemble and those with weak restraints will be very different. Regions with fewer restraints usually correspond to the flexible parts of the molecules. There is no simple measure like X-ray resolution to determine the quality of NMR structures. NMR structures can be assessed by comparing the relative differences between the models using root mean square deviation (RMSD, a structural comparison score that will be explained in Section 1.5.3), or by comparing the distribution of energetically allowed regions for backbone dihedral angles using the Ramachandran plot (Ramachandran et al., 1963).

Cryo-EM has also been used to study protein structures. Cryo-EM allows researchers to construct 3D images of microscopic objects using focused beams of electrons and super-cold temperatures. In the past, the quality of the models produced by cryo-EM was not comparable to those produced by NMR and X-ray crystallography. Recent advances - such as the development of better direct electron detectors (which give an increase in the number of images, the size of the data sets, and faster readouts), powerful computing clusters to process the large amount of data produced, improved image-processing software, and the implementation of a hybrid approach that fits high-resolution structures solved by X-ray to structures derived from cryo-EM - have improved the quality of structures tremendously (Vénien-Bryan et al., 2017). Cryo-EM does not need proteins to be crystallised and has the ability to handle large protein molecules. Recently, EM techniques have been able to solve the structure of a 465-kD protein complex (beta-galactosidase complex with a cell-permeant inhibitor) with 2.2Å resolution (Bartesaghi et al., 2015).

1.2 The structures of proteins

The functional properties of proteins depend on the 3D structure. The primary structures of proteins are the sequences of amino acids along the polymer chain. There are a total of 20 amino acids in nature. All the amino acids possess a central carbon atom to which is attached a hydrogen atom, an amino group and a carboxyl group. What differentiates one amino acid from another is the side chain (R-group) attached to the central atom.

These side chains vary considerably in their physiochemical properties. Amino acids can be grouped into three main classes: hydrophobic, charged and polar. (See Figure 1.1 for the different physiochemical properties and Figure 1.2 for the chemical formula of these amino acids). Amino acids are joined together during protein synthesis through the formation of peptide bonds (between the carboxyl and amino groups).

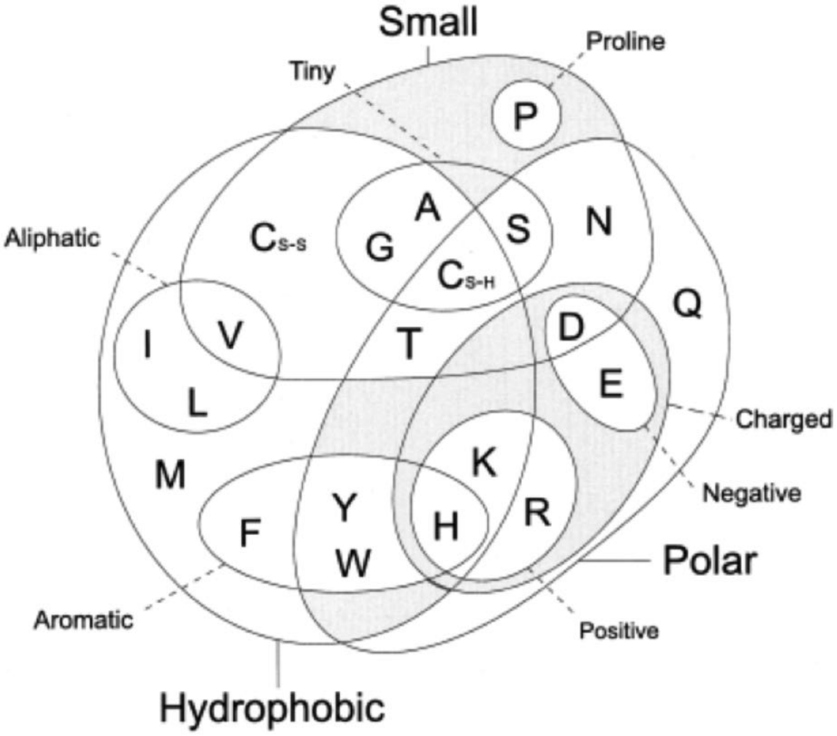


Figure 1.1: Taylor's Venn diagram (Taylor, 1986) of amino acid properties illustrating the physiochemical similarities and differences of the amino acids

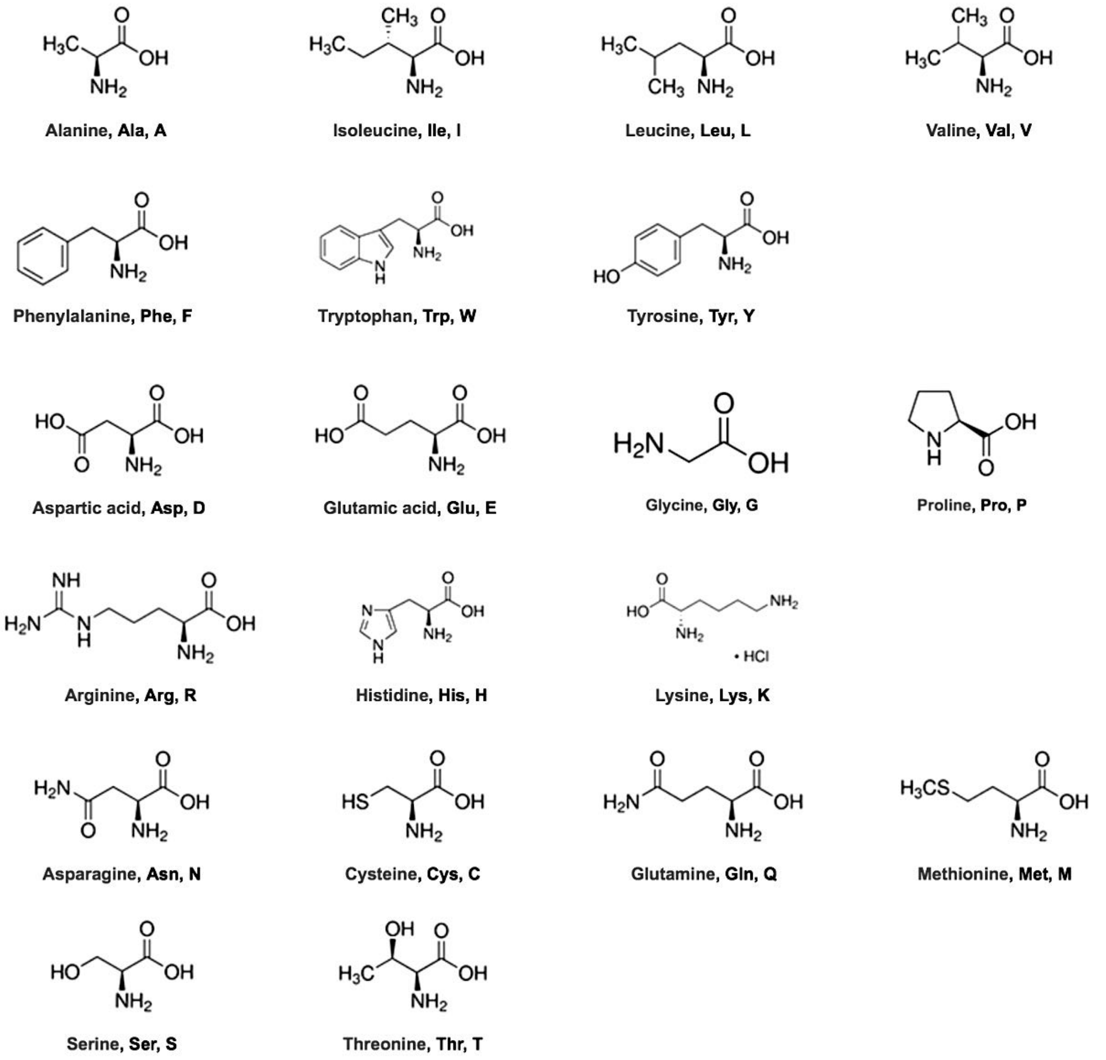


Figure 1.2: The chemical formula of the 20 amino acids

Proteins are energetically driven to fold into their 3D structure by the packing of their mainly hydrophobic amino acids into the interior, leaving a surface of hydrophilic side chains. The formation of hydrogen bonds between the backbone oxygen atoms and the amide hydrogen atoms drive different regions of the proteins to form local secondary structure elements such as α-helices and β-sheets. α-helices are the most commonly found helices in nature and usually adopt right-handed coiled rod-like structures. 3_{10} helix, π-helix, β-helix, polyproline helix, collagen helix are different configurations of helices found in proteins. β-sheets are formed by linking two or more beta strands (i.e. stretches of extended regions of the polypeptide chain typically 3 to 10 amino acids long) by hydrogen bonds and adopting sheet-like structures.

Depending on the bonding patterns between individual strands, β-sheets can adopt parallel, anti-parallel or mixed configurations.

The folding of the secondary structure elements of the proteins then gives the native 3D conformation known as the tertiary structure. The 3D conformation is usually held together by interactions between the side chains. Disulfide bridges can be formed between two cysteine residues. Ionic bonds can be formed between amino acids with an extra carboxyl group and amino acids with an extra amine group (i.e. between lysine and aspartic acids). Salt bridges are usually formed between charged side chains of acidic and basic amino acids. Hydrogen bonds can be formed between the side chain of amino acids. Hydrophobic interactions can be formed between the non-polar side chains. Some proteins adopt a quaternary structure which constitutes the assembly of different polypeptide chains via electrostatic and non-covalent bonds to give one protein complex.

Proteins that share the same common ancestor are known as homologues. Both sequence and structural similarity can be used to identify the evolutionary relationships. However, the protein structure is generally much more conserved than the protein sequence (Chothia and Lesk, 1986) and therefore very remote homologues can often only be detected by structure comparison. Homologues can be further separated into orthologues or paralogues. Orthologues are related through speciation. These proteins usually share the same protein function. Paralogues, on the other hand, occur within the same genome after gene duplication. Paralogues usually evolve new functions (i.e. modification of functional residues such as residues in the catalytic site or substrate binding site).

1.3 Aligning protein sequences

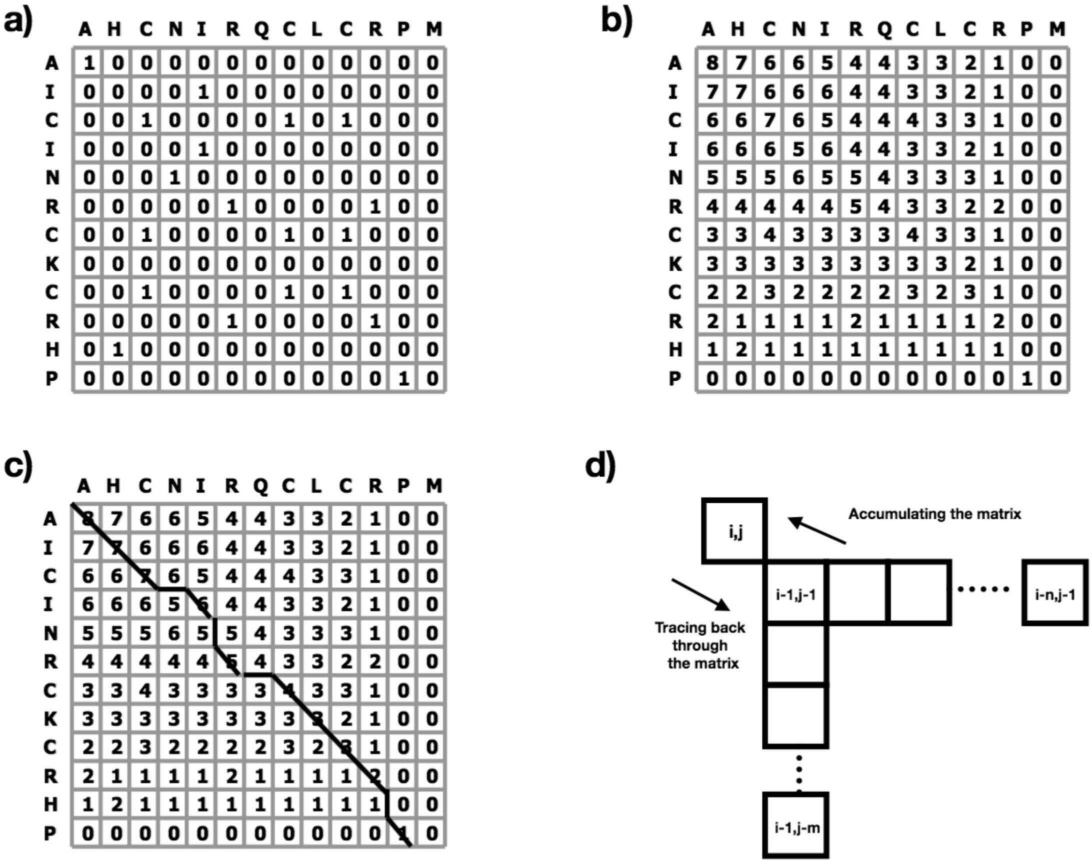
Aligning the primary sequence of a novel protein to previously characterised proteins is often important to derive information about its structure and function. Sequence alignment is the process of comparing two or more sequences in order to identify regions which are similar.

Aligning two proteins that are very similar can be straightforward. However, the process becomes complicated when diverse sequences are compared comprising multiple insertions and deletions of residues between the sequence. Gaps need to be inserted in these alignments to cope with the insertions or deletions introduced as the sequences diverged from each other. Equivalent residues need to be identified in the alignment. For example, amino acids such as leucine and valine are very similar in their physiochemical properties (both sharing comparable hydrophobicity and size). Amino acid substitution matrices such as the Identity matrix, physiochemical property matrices and evolutionary matrices are often used to score the residue matches and mismatches in alignment.

The most commonly used substitution matrices are the Dayhoff Point Accepted Mutation (PAM) matrix (Dayhoff et al., 1978) or the blocks amino acid substitution (BLOSUM) matrix (Henikoff and Henikoff, 1992). These matrices consist of all-against-all amino acid scores that reflect how often one amino acid is likely to have been mutated to the other over a given evolutionary time frame derived from an alignment of homologous protein sequences. BLOSUM matrices are derived from the amino acid changes observed in the highly conserved region (known as blocks) of >500 protein families found in the BLOCKS database (Henikoff and Henikoff, 1992). PAM matrices are based on scoring all amino acid changes likelihood in homologous sequences during evolution from 71 protein families (Dayhoff et al., 1978). The choice of which matrix to use depends on the goal of the research. BLOSUM matrices are effective to find conserved domains, whereas PAM matrices are useful to track evolutionary origins of proteins (Mount, 2008).

There are two modes of pairwise sequence alignment: global and local. Global sequence alignment attempts to align the complete sequences of the proteins to find the maximum similarity. Such methods may be appropriate when aligning sequences which are similar in length and/or multi-domain composition. In contrast, local sequence alignments attempt to find the maximum similarity within a sub-region. They are particularly useful when aligning sequences that have different lengths, or remotely related sequences that contain diverse domain regions. The Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) and the Smith-Waterman algorithm (Smith and Waterman, 1981) are used for generating global and local pairwise alignments respectively. Both of these algorithms exploit dynamic programming (DP) to produce alignment. This optimises the alignment of the identical or similar residues whist introducing the least gaps.

The first step of DP is to populate a 2D matrix with numerical scores according to the identities/similarities between the residues inside each cell (using the different substitution matrices as explained above). This is followed by the process of accumulating the scores in the 2D matrix from the bottom right corner of the matrix up to the top right corner of the matrix. Starting from the furthest right of a given row, scores are calculated one cell at a time. The score for each cell is computed by adding the score already in the cell to the best score from all the possible paths leading up to that cell (i.e. one of the cells in the row or column to the right and below the current cell). A gap penalty can also be involved to penalise the introduction of a gap in the alignment. Once the matrix has been accumulated, the highest scoring path through the matrix is traced back from top left to bottom right. Starting from the highest scoring cell in the top row, the best path will proceed through the highest scoring of the cells in a similar manner as the accumulating step. Figure 1.3 summarizes the steps in DP.



a)

	A	H	C	N	I	R	Q	C	L	C	R	P	M
A	1	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	1	0	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	1	0	1	0	0	0
I	0	0	0	0	1	0	0	0	0	0	0	0	0
N	0	0	0	1	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	1	0	0	0	0	1	0	0
C	0	0	1	0	0	0	0	1	0	1	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	1	0	1	0	0	0
R	0	0	0	0	0	1	0	0	0	0	1	0	0
H	0	1	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

b)

	A	H	C	N	I	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
I	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
I	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	5	4	3	3	3	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
H	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

c)

	A	H	C	N	I	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
I	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
I	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	5	4	3	3	3	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
H	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Figure 1.3: Steps in DP using a generic identity matrix without gap penalty. a) Scoring the matrix. b) Accumulating the matrix. c) Tracing back through the matrix. d) Mechanism for accumulating/tracing back the matrix. (Adapted from (Orengo et al., 2003))

Below are the mathematical equations for the Needleman-Wunsch algorithm (Equation 1.1) and the Smith-Waterman algorithm (Equation 1.2). Score (i,j) defines the data in a cell with coordinate i,j when comparing sequence x and y.

$$Score(i,j) = max \left\{ \begin{matrix} Score(i-1,j-1) + Score(x_i,y_j) \\ Score(i-1,j) + Gap \\ Score(i,j-1) + Gap \end{matrix} \right\} \tag{1.1}$$

$$Score(i,j) = max \left\{ \begin{matrix} Score(i-1,j-1) + Score(x_i,y_j) \\ Score(i-1,j) + Gap \\ Score(i,j-1) + Gap \\ 0 \end{matrix} \right\} \tag{1.2}$$

The sequence identity of two sequences is typically calculated as follow:

$$Sequence\ identity = \frac{Number\ of\ identical\ residue}{Number\ of\ residues\ in\ the\ smallest\ protein} \times\ 100\% \tag{1.3}$$

1.3.1 Multiple sequence alignment

Finding an optimal sequence alignment for more than three sequences using DP can be very computationally expensive and it is difficult to obtain the optimal alignment. Therefore, heuristic algorithms have been used to reduce the computational times. Progressive alignment methods build a multiple sequence alignment (MSA) by gradually combining a series of pairwise alignments starting from the most similar pair to the most distantly related ones. However, the addition of sequences sequentially to the growing MSA can be problematic as any errors made in the early stages will end up propagating to the final MSA. t-coffee (Di Tommaso et al., 2011) and Clustal omega (Sievers et al., 2011) are examples of widely used progressive alignment methods. Iterative refinement methods may refine some of the mis-alignments occurred in progressive MSA. These methods realign subgroups of sequences followed by alignment of these subgroups into a global alignment of all of the sequences. MUSCLE (Edgar, 2004) is an example of iterative based MSA program. MUSCLE first produces a rough and fast approximate MSA. The second stage of MUSCLE attempts to improve the phylogenetic tree using a better distance matrix and build a new alignment. The final stage refines the improved alignment.

MAFFT (Katoh and Standley, 2013) suite of MSA programs also adopt iterative refinement steps. MAFFT incorporates Fast Fourier transform (FFT) in which a protein sequence is converted to a sequence composed of volume and polarity values for each amino acid residue. This allows fast detection of homologous regions. Different progressive and iterative refinement methods have been implemented in various MAFFT alignment modes ranging from highly accurate for <200 sequences to very fast for >2000 sequences. Several independent benchmarks (Le et al., 2017; Pais et al., 2014; Thompson et al., 2011) have shown MAFFT outperforming other MSA methods such as MUSCLE (Edgar, 2004), t-coffee (Di Tommaso et al., 2011) and the Clustal suite of tools (Larkin et al., 2007; Sievers et al., 2011).

Generating a good MSA is key as it is possible to identify conserved residue positions from such an alignment. There are now many algorithms available for this. Scorecons is one such method that assigns a conservation score (Scorecons score) between 0 and 1 for every position in an alignment. A score of 0 indicates no conservation whilst a score of 1 indicates all residues at the position are completely conserved. The Scorecons score is calculated using amino acid similarity information from a Dayhoff-like mutation data matrix, PET1 (Jones et al., 1992). The overall score sums up the contributions from each individual sequence, and sequences are weighted inversely with their redundancy in the alignment (Valdar, 2002).

Scorecons also computes the diversity of position score (DOPS) of an alignment. DOPS reflects the proportion of diverse sequences in the alignment and the number of highly conserved positions (with high Scorecons score). An alignment with DOPS above 70 is deemed sequence diverse for analysis (i.e. having a lower probability of predicting false positives) (Dessailly et al., 2013). Conserved positions can be used to predict functional residues such as catalytic residues and substrate binding sites (Das et al., 2015; Dessailly et al., 2013). Scorecons is used in the work reported in Chapter 4 to detect whether residue changes in splice isoforms lie close to predicted functional sites.

1.4 Database searching

1.4.1 Sequence based

Generally, all of the sequence based methods involve searching for sequence matches in databases such as UniProt (UniProt, 2017) or PDB (Berman et al., 2000). BLAST (Altschul et al., 1990) is a very fast tool for scanning large sequence repositories. BLAST splits the query sequence into small fragments (with the default size of 3). These fragments are then used to search against all the fragments (available in a particular database). A substitution matrix (i.e. BLOSUM or PAM) is used to score the matches. Each fragment match is then extended in both directions, allowing for gaps, to create the largest possible segment segment pair. All the different segment pairs are scored, assigned significant value and ranked. Sequence identity and overlap between the query and segment matches are used as criteria in a typical BLAST search (Altschul et al., 1990).

1.4.2 Profile based

Sequence profiles that capture the pattern of residue types embedded in an MSA of evolutionarily related relatives can improve the sequence signal for sequence searching. These approaches have the ability to extend the sequence search into the "twilight zone" (sequence identity below 30%) and "midnight zone" (sequence identity below 20%).

Evolutionary information from families of homologous proteins can be captured in position-specific scoring matrices (PSSMs). For example, PSI- BLAST (Altschul et al., 1997) uses a PSSM to score matches between query and database sequences and is about three times more sensitive than BLAST. PSI-BLAST first employs BLAST to find the close relatives for query sequence, from which an MSA can be built. A PSSM is then generated based on the MSA and used to find new relatives. Subsequently, a new MSA and PSSM can then be built. The cycle is continued until no more new relatives can be found within a specific statistical threshold.

Hidden Markov models (HMMs) are more advanced forms of sequence profiles. HMMs were first proposed for the use in bioinformatics by Gary Churchill of Cold Spring Harbor Laboratory. The first HMM based program was SAM (Krogh et al., 1994). The revolutionary feature of HMMs is their ability to additionally capture the insertions and deletions that are found in MSAs. HMMs implement a statistical framework based on state-transition probabilities (emission probabilities and transition probabilities for moving from state to state) in an MSA. Probability is calculated for one of three states: match state (models the distribution of residues allowed), insert state (insertion of one or more residues) and delete state (deletion of one or more residues). By traversing this probabilistic network, a distribution of residues is ’emitted’ at each position to create the HMM model.

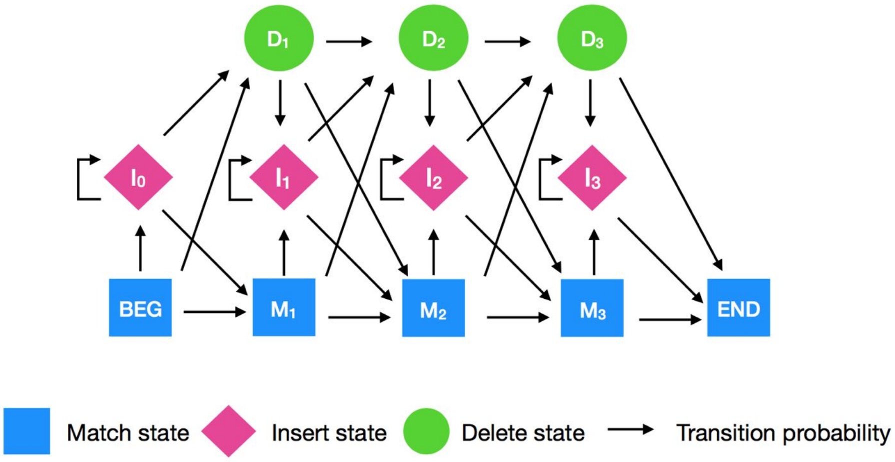


Figure 1.4: HMM model showing the transition probabilities between the different states

The HMMER (Eddy, 2011) software suite provides programs that can be used to create and manipulate profile HMMs, create databases of profile HMMs and perform sensitive searches of sequence and profile HMM databases. HMMER is widely used for making HMM models, particularly by protein family databases such as Pfam (Finn et al., 2016), SUPERFAMILY (Wilhelm et al., 2014) and Gene3D (Lam et al., 2016).

In addition, pairs of sequence profiles can be compared to find out the similarity between the two MSAs. COMPASS (Sadreyev and Grishin, 2003) generates PSSMs from two input MSAs and generates an optimal PSSM-PSSM alignment. Profile comparer (PRC) (Madera, 2008) and HHsearch (Remmert et al., 2012) supports HMM-HMM comparison and alignments.

HHsearch provides a suite of programs for HMM-based sequence searching and sequence alignment. HHsearch has been shown to perform better than HMMER and PSI-BLAST in sensitivity, alignment quality and also speed (Remmert et al., 2012). HHsearch includes predicted secondary structure information in the HMM profile. Moreover, HHsearch employs a maximum accuracy (MAC) alignment algorithm. This algorithm is inspired by the MAC algorithm introduced by Holmes and Durbin (1998). This is done by maximizing the number of correctly aligned pairs of residues (i.e. the posterior probability of match state to be aligned between 2 HMMs). HHsearch has been employed by some of the state-of-the-art 3D structural modelling servers such as Robetta (Kim et al., 2014; Ovchinnikov et al., 2017), BioSerf (Buchan et al., 2013), SWISS-MODEL (Biasini et al., 2014), nns (Joo et al., 2016) and MULTICOM (Li et al., 2015) for both template searching and sequence alignment.

1.4.3 Assessing the significance of the result obtained from these programs

When we are inferring homologies between sequences, it is important to know if a sequence match constitutes homology evidence and the likelihood that it is expected by chance. Measures such as score, bit-score and E-value are usually used to assess the quality of an alignment (match). A score S, is a numerical value that explains the overall quality of an alignment, depending on the substitution matrix and gap penalty used. Typically, the higher the score, the higher the similarity and the better the alignment. Bit score is a normalised score that allows user to estimate the magnitude of the search space. It allows the comparison of alignment irrespective of substitution matrices or gap penalty used. The bit score is calculated by:

$$Bit \quad score = \frac{\lambda S - ln(K)}{ln(2)} \tag{1.4}$$

where S is the overall alignment score, λ and K reflect the matrices and penalties used. The E-value represents the number of hits one can "expect" to see by chance when searching a database. Hence, the E-value is dependent on the size of the database and the query length. The closer the E-value to 0, the better the alignment. E-values are calculated using the following formula:

$$E = Kmne^{-\lambda S} \tag{1.5}$$

where m is the length of the query sequence, n is the total number of residues in the database, S represents the overall alignment score and K and λ are parameters dependant on the substitution matrix and on the gap penalties.

1.5 Structure comparison

1.5.1 Structure comparison methods

There are now more than 50 structure comparison methods available. Here we briefly review only those methods that are widely used by protein structure classification resources (explained in Section 1.6) or the structural modelling communities. SSAP (Taylor and Orengo, 1989) is used by the CATH resource (Dawson et al., 2017). DALI (Holm and Sander, 1993) is used by the SCOP resource (Andreeva et al., 2014). TM-align (Zhang and Skolnick, 2005) is widely used by various structural modelling groups competing in CASP (a community-wide experiment to assess the best structural prediction algorithms, more details in Chapter 2).

ProFit is based on the McLachlan rigid body superposition method (McLachlan, 1982) and uses a simple least-squares fitting algorithm which considers the optimal rotation angles of the atom-atom pairs to superpose two protein structures (Martin, 1996).

The DALI algorithm (Holm and Sander, 1993) splits the protein structures into hexapeptide fragments and then compares the contact maps of these fragments. Potentially equivalent fragments are identified by looking for similar patterns of distances between residues. These pairs are then concatenated to extend the alignment using a Monte Carlo optimisation. The structural similarity between the concatenated structures is measured after each concatenation in order to determine the quality of the alignment.

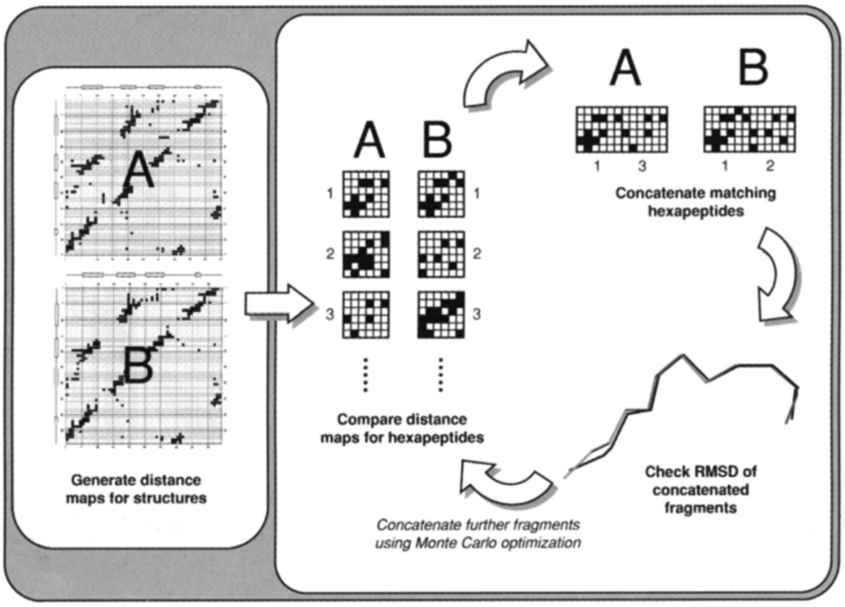

A
B
Generate distance maps for structures
A
B
1
2
3
Compare distance maps for hexapeptides
A
B
1 3
1 2
Concatenate matching hexapeptides
Check RMSD of concatenated fragments
Concatenate further fragments using Monte Carlo optimization

Figure 1.5: DALI algorithm (adapted from (Orengo et al., 2003))

Sequential structure alignment program (SSAP (Taylor and Orengo, 1989)) uses double dynamic programming (double DP) to produce a structural alignment between 2 structures based on comparing Cβ-Cβ vectors in 3D space. The program starts by determining the residue structural environments for each residue in the 2 proteins being compared. These comprise the set of vectors from each residue's Cβ atom to all the other residues' Cβ atoms in the same protein. After that, residue structural environments for potentially equivalent pairs of residues between 2 proteins are compared. Potentially equivalent pairs of residues are selected based on the secondary structure, solvent accessibility, and local conformation of the residues. Comparison of the vectors from each residue pair in the residue pair is carried out to score a 2D score matrix. DP is then used to determine the optimal path through this scoring matrix. If the residue pairs score highly enough, the optimal path is added into a 2D summary scoring matrix. This process is repeated until all potentially equivalent residue pairs are compared. Finally, another layer of DP is carried out to determine the optimal path through the summary scoring matrix, which gives the equivalent residues between the two proteins (Taylor and Orengo, 1989). Figure 1.6 illustrates the SSAP algorithm.

CATHEDRAL (Redfern et al., 2007) is the streamlined version of SSAP (Taylor and Orengo, 1989). CATHEDRAL is around 1,000 times faster than SSAP. CATHEDRAL performs an initial run of a rapid secondary structure comparison between structures using graph theory (Harrison et al., 2003). The putative fold matches identified are then aligned again using a more accurate DP similar to SSAP.

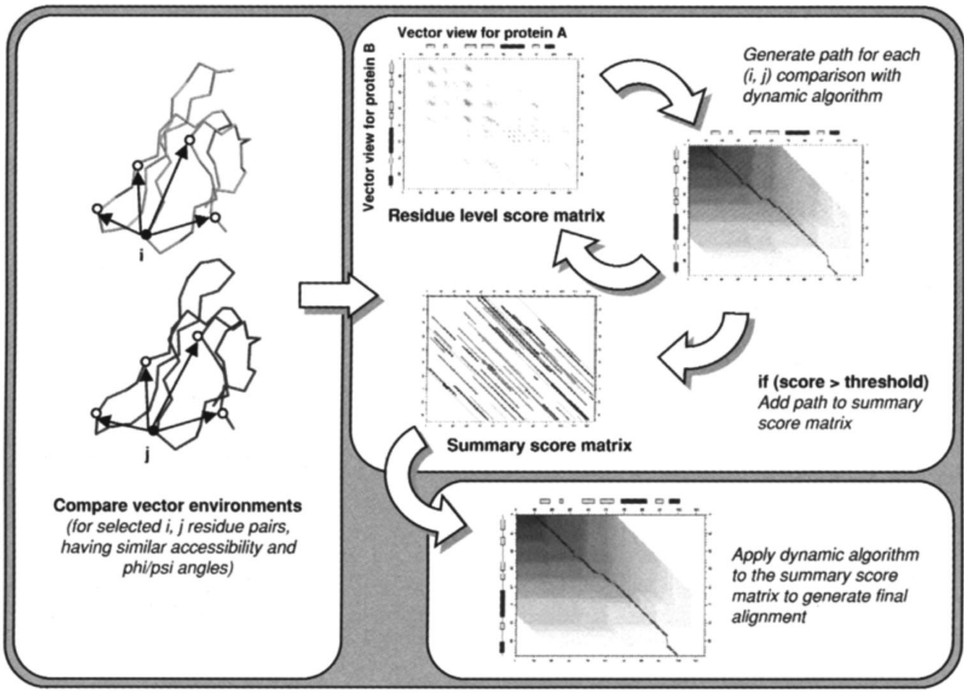

i
j
Compare vector environments
(for selected i, j residue pairs, having similar accessibility and phi/psi angles)
Vector view for protein A
Vector view for protein B
Residue level score matrix
Generate path for each (i, j) comparison with dynamic algorithm
if (score > threshold)
Add path to summary score matrix
Summary score matrix
Apply dynamic algorithm to the summary score matrix to generate final alignment

Figure 1.6: SSAP algorithm (adapted from (Orengo et al., 2003))

In 2005, Yang Zhang and Jeffrey Skolnick developed TM-align. TM-align uses three different methods to generate initial structural alignments. The first method uses DP to structurally align the secondary structures of the two proteins.

The second method performs a gapless threading of the smaller structure of the two proteins onto the larger structure (Kihara and Skolnick, 2003). Gapless threading starts by superposition of randomly chosen continuous fragments (same length in both proteins) from the longer protein onto the shorter protein. The TM-score (Zhang and Skolnick, 2004) of the superposition (explained in the next section), rotation matrix and translation factor of the superposition is recorded. This process is repeated until all possible continuous fragments are matched. The rotation matrix and translation factor from the superposition with the highest TM-score is then applied to whole protein. This produces a distance scoring matrix and DP is used to obtain the structure alignment.

The third method uses DP on a "summary" 2D matrix which combines the secondary structure scoring matrix (obtained from secondary structure similarity alignment method) and the distance scoring matrix (obtained from the gapless threading alignment method).

Based on the aligned residue information of all the initial alignments produced, a TM-align rotation matrix is produced. The structures are then superposed based on the rotation matrix. A new structural alignment can then be produced by implementing DP on a score similarity matrix (based on the residue distance between the superposed structures) with a gap-opening penalty.

Superposition of the structures is then carried out again based on the TM-align rotation matrix derived from the new alignment, followed by implementing DP on the new score similarity matrix. This heuristic iterative process is repeated until the alignment has become stable. The alignment with the highest TM-score is then returned (Zhang and Skolnick, 2005). Figure 1.7 illustrates the TM-align algorithm.

In addition, the Zhang lab also developed MM-align (Mukherjee and Zhang, 2009), an offshoot of the TM-align program designed for aligning multiple-chain protein complex structures.

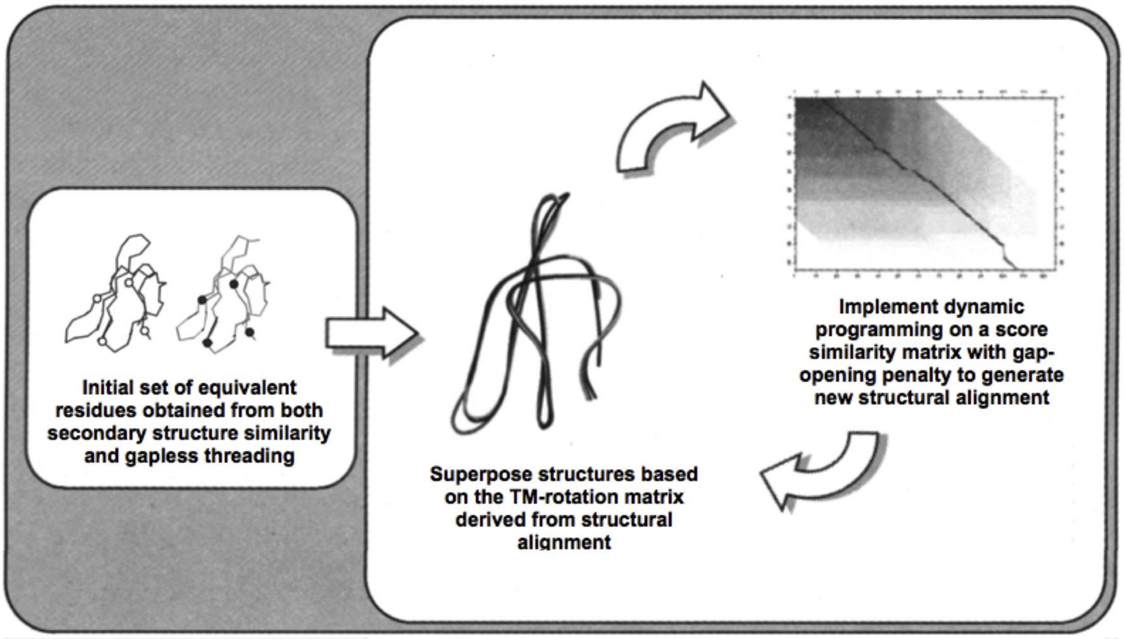

Initial set of equivalent residues obtained from both secondary structure similarity and gapless threading
Superpose structures based on the TM-rotation matrix derived from structural alignment
Implement dynamic programming on a score similarity matrix with gap-opening penalty to generate new structural alignment

Figure 1.7: The heuristic iterative algorithm of TM-align

1.5.2 Performance of structure comparison methods

In 2005, TM-align was compared against structure comparison methods CE (Shindyalov and Bourne, 1998) and DALI (Holm and Sander, 1993) using a benchmark target dataset that consists of 200 non-homologous PDB proteins. TM-align produced more accurate structural alignments and used less CPU run time compared to DALI and CE (See Table 1.1).

Methods	Coverage	RMSD	Average CPU time
CE	34.7 %	6.36 Å	2.28 s
DALI	53.5 %	14.25 Å	12.22 s
TM-align	43.4 %	5.18 Å	0.52 s

Table 1.1: Comparison between TM-align, CE and DALI (adapted from (Zhang and Skolnick, 2005))

In 2007, Redfern et al. also compared SSAP against several structural comparison methods (DALI, CE, STRUCTAL (Kolodny et al., 2005), and LSQMAN (Kleywegt and Jones, 1997)) using a dataset of 1,779 CATH-SCOP domains. The domain alignment quality of the different methods was compared. SSAP gave a higher percentage of correct folds than DALI and CE according to the SiMax scoring scheme (See Section 1.5.3 for details). SiMax takes account of the proportion of residues aligned in the larger domain structure to determine whether a significant fold match has been achieved. In addition to that, SSAP correctly aligned 25% more residues than DALI and 40% more than CE (Redfern et al., 2007).

Although no direct comparison has been done between SSAP and TM-align, both of the methods were shown to perform better than CE and DALI. Further investigation is needed to compare the two methods using the same benchmark dataset and the same scoring scheme.

1.5.3 Structure comparison scores

The most commonly used structure comparison score is the Cα root mean square deviation (RMSD). RMSD calculates the root mean square of the distance between all equivalent Cα atom pairs; Since all of the atoms are equally weighted. RMSD can be very sensitive to local changes. RMSD has a power-law dependence on the protein size; a RMSD of 3Å calculated over 20 residues is far less significant than the same RMSD calculated over 200 residues. Below is the formula used to calculate RMSD:

$$RMSD(x,y,z) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x1_i - x2_i)^2 + (y1_i - y2_i)^2 + (z1_i - z2_i)^2} \qquad (1.6)$$

To account for RMSD's over-sensitivity to local changes, Zemla et al. developed the GDT-score, which takes into account the number of Cα pairs within a distance 1Å, 2Å, 4Å, and 8Å after superposition of the two structures. GDT-TS score is calculated by taking the average of the four GDT scores obtained divided by the number of residues in the target. GDT-TS is the current official structural comparison score used in the CASP competition (which will be described the next chapter). A more sensitive GDT measure named GDT-HA is available as well, that accounts for Cα pairs which have a distance 0.5Å, 1Å, 2Å, and 4Å (Read and Chavali, 2007; Zemla, 2003; Zemla et al., 1999). The MaxSub method focuses on the percentage of Cα pairs superposing within 3.5Å distance (Siew et al., 2000). In addition, the SiMax score normalises the larger of the two proteins being compared using the following equation:

$$SiMax = max(L_1 L_2) \frac{RMSD}{Number \quad of \quad aligned \quad residues} \tag{1.7}$$

where L_1 and L_2 are the length of the respective proteins. However all these methods still fail to eliminate the dependence on protein size. In 2004, Zhang and Skolnick proposed a solution to the problem by introducing a protein size-dependent scale and developed the TM-score:

$$TM - score = Max \left[\frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \left(\frac{d_i}{d_0} \right)^2} \right] \tag{1.8}$$

where L_N is the length of the template structure, L_T is the number of the residues aligned to the template structure, d_i is the distance between the i-th pair of aligned residues and d_0 is the scaling factor. The scaling factor is calculated as $d_0 = 1.24\sqrt[3]{L_Q - 15} - 1.8$, where L_Q denotes the length of the model.

Xu and Zhang have suggested that the TM-score has certain advantages in scoring structure similarities over other methods such as MaxSub, RMSD, and GDT-TS. It is thought to be more sensitive to global topology, accounts for all residues pairs and has been used as a quantitative criterion for protein topology classification. Protein pairs with TM-score above 0.5 are generally in the same CATH/SCOP topology fold (Xu and Zhang, 2010).

Other scores include the SSAP score which has been used for protein structure classification in the CATH database (Dawson et al., 2017). SSAP score measures the vectors from all the equivalent residues pairs of a protein to other equivalent residue pairs in the protein, followed by the normalization of the score (to be independent of protein size) using the following formula:

$$S_{SSAP} = \left(\sum_{i=1,i'=1}^{aln} \sum_{j=1,j'=1}^{aln} Svector_{i \to j, i' \to j'} \right) / (maxequivs^{*}(maxequivs - 11)) \quad (1.9)$$

where aln is the number of aligned residues, maxequivs is the maximum number of equivalent positions between proteins. The vector similarity score $Svector$ is calculated as follows:

$$Svector_{i \to j, i' \to j'} = \frac{500}{(10 + \delta)} \quad (1.10)$$

where δ is distance between vectors. The SSAP score has a maximum of 50. To set the final SSAP score into a maximum of 100, the following equation is used:

$$S'_{SSAP} = ln(S_{SSAP}) \times \frac{100}{ln(50)} \quad (1.11)$$

A SSAP score of 70 out of 100 indicates structures have a similar fold (Taylor and Orengo, 1989).

1.6 Protein structure classification

Protein structures have been organised into evolutionary families by resources such as CATH (Dawson et al., 2017) and SCOP (Andreeva et al., 2014). Both of these resources classified the proteins at the domain level. Protein domains are independently-folding functional and evolutionary units of a protein sequence. They are generally observed as local, compact units of structure, with a hydrophobic interior and a hydrophilic exterior, forming a globular-like state that cannot be further subdivided. CATH and SCOP identify structural domain boundaries within protein structures and classify domains into protein families based on their evolutionary origin. However, there are differences between the two resources (i.e. how the proteins are chopped into domains and how the domains are classified into families) which will be discussed further below.

Sequence-based information has also been used to classify proteins into protein families. Resources employing these techniques usually identify conserved sequence patterns. The most widely used resource is Pfam (Finn et al., 2016). The newest release of Pfam (release 31) comprises 16,712 protein families, which covers more than 75% of the sequences in UniProt.

1.6.1 CATH

CATH (Orengo et al., 1997) was established by Orengo and Thornton in the mid-1990s. The first release of CATH in 1997 featured a total of 8,000 structural domains classified into 1,000 superfamilies. The newest release of CATH (CATH version 4.1) classified 308,999 structural domains into 2,737 superfamilies.

CATH is derived using a semi-automated approach followed by expert manual curation. Domain boundaries are inferred in new protein structures by searching for using both sequence and structural similarity to domains that have been classified in CATH. Sequence similarity is obtained by scanning the new chains against a library of CATH superfamily HMMs. Structural similarity is computed using a streamlined version of SSAP (Taylor and Orengo, 1989), CATHEDRAL (Redfern et al., 2007), that is around 1,000 times faster. Both of these similarities are then used to assess whether the new chain is homologous to any domains in CATH. For cases with low similarity, the domain boundaries are identified manually. Manual assignment of domain boundaries is guided by three *ab initio* domain identification methods: PUU (Holm and Sander, 1994), DETECTIVE (Swindells, 1995) and DOMAK (Siddiqui and Barton, 1995).

The CATH superfamily code is denoted by four numbers corresponding to each level in the CATH classification (i.e. 3.20.20.120). At the top of the hierarchy is the class level where structural domains are classified based on their secondary structure content (i.e. the proportion of residues adopting α-helical or β-sheet conformations). Class 1 domains are mostly α-helical, Class 2 domains are mostly β-sheet, Class 3 domains have a significant amount of both α-helical and β-sheet. Class 4 domains have very little secondary structure content.

The second level of the hierarchy is the architecture level given by the global arrangement of secondary structures in 3D space. This is followed by the topology level where domains with similar folds (which takes into account the 3D arrangement, orientations and connections between the secondary structures) are grouped together. The fourth level is the homologous superfamily level where domains are deemed homologous if they meet two of the following three criteria:

- Significant sequence similarity (e.g. Significant E-value)
- Significant structural similarity (e.g. SSAP score above 80 with at least 60% overlapping residues)
- Evidence of functional similarity (e.g. conserved catalytic, or binding site residues and cofactors)

1.6.1.1 Gene3D

For each CATH superfamily, HMMs are built for non redundant representatives (at 95% sequence identity). Based on these HMMs, the CATH-Gene3D resource (Lam et al., 2016) predicts domain superfamily assignments in UniProt sequences (UniProt, 2017) and ENSEMBL sequences (Aken et al., 2016). These annotations provide structural and functional insights to the protein. The current release of Gene3D covers a total of 54 millions sequences belonging to 2,737 CATH superfamilies, from 19,000 genomes.

In addition to the domain assignments, Gene3D also provides protein-protein interaction information, drug information, catalytic site information and mutation information for human proteins. For regions that are not covered by structural domains, Gene3D also predicts if these regions are intrinsically disordered using IUPRED (Dosztanyi et al., 2005). A total of 60,000 3D structural models were built for over 70% of the structurally uncharacterised sequences in human and fly genomes using an in-house modelling platform FunMod described in Chapter 2.

1.6.1.2 Sub-classification of homologous superfamilies into functional families (FunFams)

Once the sequence relatives are assigned to a specific superfamily in CATH, functional families are identified by performing agglomerative clustering of domain relatives (GEMMA (Lee et al., 2010)). Relatives are first annotated with Gene Ontology (GO) annotations and then clustered at 90% sequence identity (S90). S90 clusters without GO-annotations are removed. In addition, sequence fragments (less than 80% of the average sequence length) are also removed from the clusters. For each S90 clusters, a sequence profile is built based on the MSA. S90 clusters are then compared using HMM-HMM profile comparison software, COMPASS (Sadreyev and Grishin, 2003). At each round, clusters that match a specific threshold are merged and profiles are generated for the new clusters. The process is continued to until a single cluster remains (i.e. producing a hierarchical clustering tree from the leaf nodes to the root).

Determining an optimal cut of a hierarchical clustering tree of sequence relatives within a given superfamily is key to producing functionally pure clusters. Initially, GO annotations were used to ensure the functional coherence in each functional family (i.e. clusters are only merged if they contain coherent GO terms) (Rentzsch and Orengo, 2013). However, due to the paucity of the GO terms and annotation biases existing in the GO (Schnoes et al., 2013), a newer approach that exploits sequence patterns and unaffected by the limitations of GO was developed.

The new FunFHMMer protocol determines the best optimal cut of the tree using conserved positions and specificity determining positions in the cluster alignments (Das et al., 2015). Highly conserved positions are generally important for the stability, folding or function of the protein domain. Specificity-determining positions are positions that are conserved within and unique to a particular cluster, sharing a specific function and usually involved in functional divergence from other clusters (Abhiman and Sonnhammer, 2005; Rausell et al., 2010).

The functional purity of the new FunFams was demonstrated in a number of ways:

by validating against experimentally determined Enzyme Commission (EC) (Webb, 1992) and SFLD (Akiva et al., 2014) annotations and also by checking whether known functional sites coincide with highly conserved residues in the MSAs of FunFams (Das et al., 2015). Functional predictions based on FunFams were ranked amongst the top five methods for the "Molecular Function" category and the "Biological Process" category in the 2nd CAFA International Function Prediction experiment (Jiang et al., 2016).

1.6.2 SCOP

The Structural Classification Of Proteins (SCOP) database (Murzin et al., 1995), like CATH, classifies structural domains hierarchically into three different levels. SCOP is largely based on expert curation resulting in a very high-quality dataset. The highest level of classification in SCOP is the Class level which is based on secondary structure content. The structural domains are further classified based on their fold groups, which is similar to the T-level in CATH. The third level of SCOP is the homologous superfamilies level (comparable to the H-level in CATH) which comprises of either domains which share sequence identity with at least 30% sequence identity or those having very similar structures and functions. SCOP uses tools such as the structural comparison tool DALI (Holm and Sander, 1993), Pfam (Finn et al., 2016) and BLAST (Camacho et al., 2009) to check for the homology of the domains (Andreeva et al., 2014).

Recently, a new prototype of SCOP, SCOP2 (Andreeva et al., 2014) has been introduced to organise the structural domains. With the expansion of structural data in PDB, the SCOP team recognised that evolutionary relationships can be more complex than first thought, leading to a new design to capture the relationships. Rather than using a hierarchy, the new SCOP2 uses a directed acyclic graph to represent the structural and evolutionary relationship of the domains. Both structural and evolutionary relationships are now split into two separate categories allowing the classification of the homologous proteins into different folds and structural classes while keeping them in the same evolutionary family and superfamily (Andreeva et al., 2014).

SCOPe (Structural Classification of Proteins-extended) (Fox et al., 2013) is a database that extends the SCOP v1.75 database (the newest stable release of SCOP). To keep up with structural growth in PDB, SCOPe uses automated methods to classify the new structures using the original SCOP-based hierarchy. Similar to Gene3D, the SUPERFAMILY (Wilson et al., 2009) resource provides domain superfamily and family assignments for protein sequences based on SCOP.

1.7 Overview of Thesis

The first work chapter of the thesis explores the structural coherence of domains in FunFams (built using the new FunFHMMer algorithm). We then explore the use of FunFams in protein monomer modelling and a new modelling pipeline (FunMod) that, utilises the FunFam protocol to model structurally uncharacterised domain sequences from human and fly genomes.

In the second work chapter, we explore the use of FunFams in protein complex modelling. We report analyses that explored the conservation of domain-domain interfaces in FunFams. A new template library is generated for template selection in 3D-template based docking and we report a FunFam based complex modelling protocol.

In the final work chapter, we employ the FunFam based structural modelling tool to understand the implications of alternative splicing. We focused on mutually exclusive exons (MXEs) for which there is more evidence of translation into proteins from proteomic experiments. Our pipeline allowed us to assess for the first time, structural and functional consequences of MXE events in whole genomes.

Chapter 2

Modelling protein monomers

2.1 Introduction

Proteins are essential macromolecules found in all biological systems. They are involved in virtually every biological process. For instance, the cooperation of transcription factors, activators and DNA-binding proteins are necessary for the mRNA transcription process that ultimately drives the production of proteins.

The 3D atomic coordinates of protein structures can be solved using experimental methods such as X-ray, NMR or electron microscopy. Once a protein structure is determined, it is usually deposited in the Protein Data Bank (PDB) (Berman et al., 2000). The PDB is the most comprehensive protein structure repository containing over 130,000 entries. However, resolving the protein structure is complicated and time-consuming, therefore the number of solved protein structures is much less than the number of proteins sequenced so far, which is $\geq$89 million (UniProt, 2017). This gap is known as the "Protein Structure Gap". Various computational protein structure modelling approaches have been developed to bridge the gap and will be described in the next section.

2.1.1 Predicting protein structures

The most commonly used and most accurate protein structure modelling method is comparative modelling which predicts the structure of an unknown protein using known information from one or more homologous partners. Comparative modelling usually involves three steps: (1) Identification of template structures for modelling the query protein, (2) Sequence alignment between the template and the query, and (3) Modelling the structure of the query.

Generally, all the template selection methods involve searching for template protein structures from the PDB. Global sequence identity between the query and templates has been used extensively as the primary criterion for a search process using BLAST (Altschul et al., 1990). Comparative modelling of a structurally uncharacterised protein generally produces a good 3D model as long as a homologous template with a global sequence identity of 30% or more is used. However, once the sequence identity falls below 30% (Twilight zone), the model quality deteriorates rapidly (Baker and Sali, 2001).

There have been some powerful methods that have been developed to recognise remotely related structural templates such as PSI-BLAST (Altschul et al., 1997), which is based on 1D sequence profiles, and HHsearch (Söding, 2005) and HMMER (Eddy, 2011), which are based on HMMs. These approaches have the ability to extend the sequence search into the "Twilight Zone" and find templates which have high structural similarity with the query. Phyre2 (Kelley et al., 2015), MULTICOM (Cao et al., 2015) and HHpred (Söding et al., 2005), are examples of robust comparative modeling servers that use HHsearch to search for structural templates.

Various studies have highlighted the importance of considering the physical and structural environment of the template selected for modelling a particular query sequence such as pH, temperature, space group and quaternary structures (Fiser, 2004). However, Sadowski and Jones concluded that these factors do not significantly improve template selection for single domain modelling (2007). If there is more than one potential template with comparable sequence identity, it is preferable to use the template with the best X-ray resolution, regardless of conditions.

Once a structural template has been identified, both the template and alignment (usually obtained from the template searching method) are submitted to a comparative modelling program that will predict the 3D atomic coordinates of the query protein. Overall, it is generally agreed that profile-based alignments produce better quality models than sequence-based alignments (Yan et al., 2013). In addition, HMM-based alignments (i.e. HHsearch) tend to give higher quality models than PSSM-based alignments (i.e. PSI-BLAST) (Yan et al., 2013).

The major steps in modelling the structure of a query sequence, based on a template structure, are summarised below. For a more detailed account see recent reviews (Mishra et al., 2013; Tress, 2013). Guided by the sequence-template alignment, comparative modelling methods usually start by copying the coordinates (structurally conserved regions) from the template to assemble the basic backbone of the model.

Processing deleted residues between the query and template sequence involves the removal of residues and closure of the hole formed by creating the new peptide-bond. For insertions, loop modelling can be done by searching through high resolution fragment libraries (either derived from the PDB or structural domain resources such as CATH or SCOP) to find segments that fit the specific part of the backbone. However, these methods are limited by the fact that the number of possible conformations increases exponentially with the length of a loop (difficult when the loop size is $>$7 residues). By contrast, conformational approaches construct loops by searching through the conformational space of possible loop conformations driven by satisfying a specific energy function (e.g. stereo-chemical, distance, steric constraints). In order to maximize the accuracy of loop prediction, simulating the correct environment (energy functions) is key. Approaches to do this include hybrid methods which employ both knowledge based and physics-based energy functions (see for example Park et al. 2014 for more details) and physics-based energy functions, such as CHARMM36m (Huang et al., 2017).

The next step is side chain modelling which involves the process of refining/adding side chains to the backbone built. Strategies such as dead end elimination, Monte Carlo sampling and simulated annealing are usually used to sample the most probable rotamer (side chain conformation), based on the local conformation of the backbone, from rotamer libraries such as SCWRL4 (Krivov et al., 2009). Once the model is produced, it is usually refined to minimize unfavourable collisions between atoms. This is usually done by performing energy minimizations following molecular dynamics simulations using force fields. Excessive refinement may cause the model to be deviate significantly from the original template (For some recent approaches see Feig 2016; Kim and Kihara 2016; Lee et al. 2016; Park et al. 2016).

In 1993, Andrej Sali and Tom Blundell developed MODELLER, which is still one of the most widely used comparative modelling methods. It is based on satisfaction of spatial restraints theory inspired by NMR spectroscopy. These restraints include homology-derived constraints obtained from the alignment of query sequences and template structures, stereochemical restraints extracted from the CHARMM-22 molecular force field (Brooks et al., 1983), and statistical restraints compiled from a list of known protein structures. Based on the alignment between the query and the model, a set of spatial restraints are derived. These restraints which include bond distances, bond angles, dihedral angles, Van der Waals repulsion, are then expressed as probability density functions. These functions are then combined into an objective function used to calculate the location of each atom in the protein (Sali and Blundell, 1993).

Besides MODELLER, there are other comparative modelling programs available such as nest (Petrey et al., 2003), 3D-JIGSAW (Bates et al., 2001), Builder (Koehl and Delarue, 1995), SWISS-MODEL (Kopp and Schwede, 2004), and SegMod/ENCAD (Levitt, 1983, 1992). A comparison was done between all these methods by Wallner and Elofsson (2005). This showed that no single program outperformed the others. All of them have their advantages and disadvantages. But, MODELLER, nest, and SegMod/ENCAD produce more chemically correct models than the other programs.

Following the introduction of MODELLER, many other approaches were developed for protein structure prediction. To assess their performance and identify which features work best, an independent assessment initiative was established in 1994 (Moult et al., 1995). The Critical Assessment of Protein Structure Prediction (CASP) is a community-wide experiment that is held bi-annually. Whilst CASP1 had only 3 categories (comparative modelling, fold recognition and *ab initio* modelling), many more categories have been introduced since then, such as accuracy of predictions for residue-residue contacts and disordered regions. Other categories include model quality assessment, model refinement, data-assisted prediction, protein complex prediction, and recently, prediction of biological relevance. All these categories are important in structural modelling (Moult et al., 2016), and a few of them are highlighted below, particularly those relating to recent developments in comparative modelling.

2.1.2 Recent developments in structural modelling

An exciting recent development relates to more accurate predictions for residue-residue contacts. Residue contact information has been used in the past, albeit not very successfully (i.e. with $>$80% of false positives, (Monastyrskyy et al., 2014)), and whilst these approaches included co-evolution methods, the performance was poor because it was difficult to separate indirect couplings from direct couplings. In addition, very sequence diverse MSAs were typically required. Recently, methods based on direct coupling analysis have been able to disentangle direct couplings from indirect couplings (Jones et al., 2012; Kamisetty et al., 2013; Marks et al., 2011; Nugent and Jones, 2012). Furthermore, for some cases the problem of obtaining a sufficient number of diverse sequences can be solved by using metagenome data (Ovchinnikov et al., 2017).

In addition, machine learning approaches (recently deep learning), that utilise features related to the residue type (i.e. polarity etc.), structural characteristics (i.e. solvent exposure, secondary structure etc.), sequence separation length between the residues under consideration, and pairwise information between all the residues involved, also show promise in contact prediction (Adhikari and Cheng, 2016; Eickholt and Cheng, 2012; Feinauer et al., 2014).The best residue contact predictor in CASP11 (Monastyrskyy et al., 2015) was MetaPSICOV (Jones et al., 2015; Kosciolek and Jones, 2016), which integrates both co-evolution and machine learning methods. Since then, many more structural groups have started to employ residue contacts using integrative methods (He et al., 2017; Skwark et al., 2014) or deep learning methods (Wang et al., 2017) ultimately using this data to guide 3D structure modelling. In the template free category of CASP11, an accurate structural model of a 256 residue protein was successfully generated by incorporating contact information (Monastyrskyy et al., 2015). In addition, residue contact data can be used for model ranking, selection, evaluation and refinement (Adhikari and Cheng, 2016; Park et al., 2016).

Other recent developments are the application of different profile based methods in template identification and sequence alignment (Markov random fields, (Ma et al., 2014), conditional random forests, (Joo et al., 2016)); the use of integrated template-based and *ab initio* approaches (Yang et al., 2016); better protein model refinement methods with improved energy functions and MD simulations (Della Corte et al., 2016; Feig, 2016; Kim and Kihara, 2016; Lee et al., 2016; Park et al., 2016).

However, merely producing a model without determining the quality of the model is unreliable if the structure is to be used to derive biological insights. The chapter continues with the explanation on how to assess the quality of protein models.

2.1.3 Assessing protein models

If there is a native structure for the protein, then the best assessment of model quality is by comparison with the known structure. Structural comparison is usually done by superposing the 3D model against the native protein structure. This is the most direct way of determining how well the model captures the characteristics of the native structure. Different structural comparison methods and scores have been developed and are described in Chapter 1.

However, if there is no native structure available, other model quality assessment approaches must be used. The performance of these approaches is also usually assessed by doing a structural comparison to known protein structures. These approaches will be described in the following section.

To assess the performance of a protein modelling algorithm, a benchmark set of proteins is used for which the structure is already known. Model quality can then be assessed by comparing the 3D model against the known structure.

2.1.3.1 Protein monomer model quality assessment

A good quality protein model should resemble a native protein. Native proteins usually have compact, well-packed 3D-structures. The spatial features of the residues should comply with empirically characterised constraints on torsional angles captured in Ramachandran plots (Ramachandran et al., 1963). Hydrophobic side chains of the protein are buried to reduce unfavourable contacts with water molecules. Hydrogen bonds, disulphide bridges, salt bridges and covalent bonds should be present, as these facilitate the folding and packing of the polypeptide chain.

These methods typically used by structural biologists to check whether their crystal structures are well determined include PROCHECK (Laskowski et al., 1993) and MolProbity (Chen et al., 2009) which determine whether a protein structure has native-like features. These methods use various approaches to rule out unlikely protein structures with unfavourable stereochemical properties such as Ramachandran outliers, steric clashes, incorrect hydrogen bonds and distorted bond angles.

From a thermodynamic perspective, native proteins are usually folded to the lowest energy state (Rangwala and Karypis, 2010). Many energy-based programs have been developed to select the most native-like model, with the lowest energy state, from decoy sets. Statistical potential energy based functions are derived from the statistical analysis of the growing numbers of experimental protein structures. In contrast, physics-based energy functions use molecular mechanics force fields of molecules that take into account bond lengths, torsional angles, Van der Waal forces and electrostatic interactions (Brooks et al., 1983; Weiner et al., 1984).

In addition, the quality of protein models can also be assessed by checking the compatibility of the models produced with the conservation of the sequence pattern. The core of the proteins is usually composed of conserved residues. In contrast, protein surface residues tend to be less conserved with more variability (Branden and Tooze, 1999).

The current state-of-the-art model quality assessment methods can be divided into two main types: single model methods and clustering methods.

Single model methods Single model methods use evolutionary information (Kalman and Ben-Tal, 2010), statistical potentials, knowledge-based potentials (Brooks et al., 1983; Weiner et al., 1984) and a combination of different features (Benkert et al., 2008; Cao et al., 2016; Liu et al., 2016; Singh et al., 2016) obtained from only one model to evaluate the model quality (Wallner and Elofsson, 2003).

The most commonly used statistical potential based model quality assessment method is MODELLER's DOPE (Shen and Sali, 2006). DOPE is an atomic distance-dependent statistical potential based on a physical reference state that accounts for the finite size and spherical shape of proteins. There are many statistical potential methods available. They differ in the sample set of known protein structures used, the protein representation (e.g. all atoms, Cα atoms), the spatial features (e.g. angles, distances, solvent accessibility), and the definition of the reference state (Dong et al., 2013). Recently, new methods such as GOAP (Zhou and Skolnick, 2011), SOAP (Dong et al., 2013), DOOP (Chae et al., 2015) and VoroMQA (Olechnovič and Venclovas, 2017) have been introduced and all claimed to be more reliable than their counterparts.

Model quality assessment methods exploiting machine learning methods are also becoming popular. The major advantage of ML methods is their ability to take into account a large number of features simultaneously, often capturing the hidden relationships among them, which is hard to deduce with energy term measures alone. ProQ2 combines evolutionary information, MSA data, and structural features from the model using a support vector machine (SVM) to assess the quality (Ray et al., 2012). The recent ProQ3 uses a deep learning method to combine ProQ2 with Rosetta energy terms (Leaver-Fay et al., 2011) and has been shown to be superior to ProQ2 (Uziela et al., 2017, 2016). DeepQA is another deep learning method that combines physiochemical properties (i.e. secondary structure similarity, solvent accessibility) and statistical potential energy terms (Cao et al., 2016). MQAPRank is a machine-learning-to-rank method that extracts features from statistical potentials and the scores obtained from a few model quality assessment methods (Jing et al., 2016).

SVMQA is an SVM method that combines 8 statistical potential energy terms and 11 consistency-based terms (between the predicted values from the sequence of the query protein and calculated values from the model built) (Manavalan and Lee, 2017).

Besides assessing the model from a global perspective, local quality assessments of protein models are also available. It is possible to discriminate between good/bad-modelled regions of a whole protein chain using software such as QMEAN (Benkert et al., 2008), ProQ2 (Ray et al., 2012), and ModFOLD (Maghrabi and McGuffin, 2017).

Clustering methods In contrast to single model methods, clustering methods are based on the structural comparison of multiple models generated for a single target. All-against-all structural comparisons are first carried out and the resulting scores used to generate a N-dimensional distance matrix based on the structural distances between each model (See Figure 2.1).

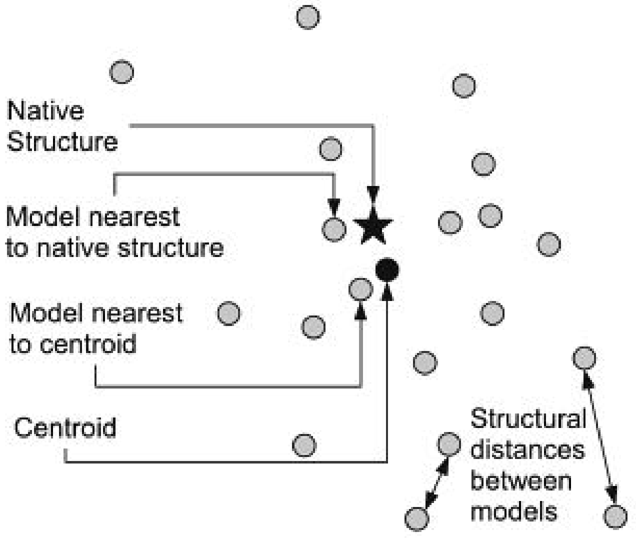

Native Structure
Model nearest to native structure
Model nearest to centroid
Centroid
Structural distances between models

Figure 2.1: Clustering of models according to structural distance from each other (adapted from Rangwala et al. 2012).

These approaches assume that the best model is the model structure with the lowest average distance to the rest of the dataset (Konopka et al., 2012). Therefore, after clustering the models, these approaches select the centroid for each cluster. The best model of the whole decoy dataset usually lies within the largest structurally conserved cluster. A model quality score for the model is calculated by averaging the structural comparison scores obtained from all the pairwise comparisons (model vs model) within the cluster and is usually followed by normalisation of the score. Recent methods that use clustering approaches are PconsD (Skwark and Elofsson, 2013), MULTICOM-CONSTRUCT (Cao et al., 2014) and ModFOLD6-rank/ModFOLD6-cor (Maghrabi and McGuffin, 2017).

2.1.3.2 Recent developments in model quality assessment

Model quality assessment by clustering has typically been superior to other quality assessment methods. However, these approaches fail to identify good quality models if the majority of the models are of bad quality and are structurally similar to each other. The other problem with clustering methods is the high computational costs.

Furthermore, many single model methods have recently been identified that can achieve better performance than clustering methods, for example, in the CASP category that selects the good-quality models from decoys (please refer to the CASP12 quality assessment result page, http://predictioncenter.org/casp12/qa_diff2best.cgi). This is probably due to rise of machine learning methods. SVMQA is an SVM method that is based on the combination of two independent predictors trained on the TM-score or GDT-TS score (Manavalan and Lee, 2017). Other methods exploit deep learning and machine-to-rank which seem to be superior to SVMs (Cao et al., 2016; Jing et al., 2016; Uziela et al., 2017, 2016).

2.1.4 Resources dedicated to large-scale comparative modelling of genome sequences

As mentioned already above, there have been several recent developments in comparative modelling and many excellent servers are now available for biologists wishing to model the structure of a query protein (see (Modi et al., 2016) for more information on the servers that are currently highly ranked). Since the focus in this chapter is more related to providing libraries of structural templates and a library of structural models, we consider resources providing large repositories of pre-calculated 3D models.

In particular, four resources that provide pre-calculated 3D structural models for over 100,000 UniProt sequences (for multiple model organisms) are described below and for each we describe how the structural models are built. These resources provide easy access to 3D structure data, visualize these structures using state-of-the-art visualization platforms and also provide functional annotations, where available e.g. inherited binding site information and other information valuable for life science researchers.

2.1.4.1 ModBase

ModBase (Pierce et al., 2014) was developed by the Sali group in 1998 and currently contains more than 36,000,000 protein models (5,956,279 unique sequences) for at least 66 species (as of April 2017). $\sim$82% of the 170,418 human transcripts in the database are annotated with structural models. ModBase uses ModPipe (Eswar et al., 2003), an automated pipeline, to produce the models. ModPipe utilizes a whole range of template selection methods (sequence-sequence, sequence-profile, profile-profile) including PSI-BLAST and HHsearch. The alignment obtained from the template selection method is then fed into MODELLER for the modelling process. For each model ModBase provides five different quality assessment criteria (sequence identity, GA341 (Melo et al., 2002), normalised DOPE (Shen and Sali, 2006), ModPipe Quality Score, TSVMod score (Eramian et al., 2008)).

In addition to the model quality, the target-template alignment and sequence identity are also provided. In addition, some of the entries contain information about putative ligand binding sites, SNP annotation, and protein-protein interactions.

2.1.4.2 SWISS-MODEL repository

SWISS-MODEL (Bienert et al., 2017) is another comprehensive repository providing 3D structural models for the 12 most accessed genomes in UniProtKB. It houses more than 900,000 models for UniProt sequences. Out of the 21,042 human sequences,

$\sim$75% are annotated with at least 1 structural model. SWISS-MODEL repository also provides structural models for homo-oligomeric complexes. All the homology models were created using an in-house modelling platform PROMOD3 (Bienert et al., 2017) that uses BLAST and HHsearch for template searching. In order to facilitate oligomeric complex modelling, structural templates in the database are also organised as quaternary structure assemblies. The database is updated weekly and contains more than $\sim$81,000 unique sequences in $\sim$180,000 assemblies. QMEAN (Benkert et al., 2008) is used to assess the quality of the models. As well as model quality, all models are provided with the target-template alignment and sequence identity. Some of the entries contain InterPro functional annotations (Finn et al., 2017). SWISS-MODEL plans to model more homo-oligomeric complexes, even for distant relatives, and to possibly include hetero-oligomeric complexes.

2.1.4.3 Protein Model Portal

Protein Model Portal (Haas et al., 2013) is a database which collects both experimental structures and structural models. As well as structural models found in the ModBase and SWISS-MODEL repositories, models generated by some of the NIH funded Protein Structure Initiative (PSI) centres are also included. Based on the UniProt release 2017-1, the portal comprises 5,388,221 unique sequences covered by at least one model. By combining models from different resources, the suppliers of the Protein Model Portal can apply the same model quality assessment and validation criteria to them. Again, each model is provided with the sequence-template alignment and sequence identity. The user can also request further assessment of model quality as the portal provides a submission interface to other quality assessment servers such as ModEval (Eramian et al., 2008), QMEAN (Benkert et al., 2009) and ModFOLD (Maghrabi and McGuffin, 2017). Furthermore, the models provided by different resources can be structurally superposed to analyse the variability amongst them. For any queries with no structural model currently available, the portal provides a submission interface to modelling servers such as I-TASSER (Yang et al., 2015), Phyre2

(Kelley et al., 2015).

2.1.4.4 Genome3D initiative

Genome3D (Lewis et al., 2015) is a UK-based collaborative project to annotate genome sequences with structural information. The participating partners include Gene3D (Lam et al., 2016), SUPERFAMILY (Wilson et al., 2009), Phyre2 (Kelley et al., 2015), VIVACE (Ochoa-Montaño et al., 2015), pDomTHREADER (Lobley et al., 2009) and BioSerf (Buchan et al., 2013). Each resource provides models based on either SCOP or CATH domain structures. Therefore, to facilitate the comparison of predicted models, Genome3D identifies matching CATH and SCOP superfamily pairs. Genome3D uses both homology-based approaches (e.g. Gene3D, SUPERFAMILY, Phyre2) and threading-based approaches (e.g. FUGUE, pDomTHREADER, Phyre2) to provide structural annotations for UniProt sequences. Genome3D annotates 94.6% of the 20,195 human sequences with at least one structural domain annotation. In addition to that, 88% of the 20,195 human sequences are annotated with 3D structural models. Structural models in the resource were built by the following comparative modelling and threading methods.

BioSerf (Buchan et al., 2013) is a fully automated pipeline that combines comparative modelling, protein threading and *ab initio* approaches. BioSerf searches for a suitable homologous template using PSI-BLAST and HHsearch. MODELLER is then used to build the model. Protein threading is done using in-house threading methods pGenTHREADER (Lobley et al., 2009) and pDomTHREADER (Lobley et al., 2009) guided by a protein secondary structure prediction method PSIPRED (Jones, 1999). The FRAGFOLD algorithm is used, where appropriate, to create *ab initio* models. FRAGFOLD uses known protein super secondary structural fragments and uses a simulated-annealing algorithm to assemble the most probable 3D protein structure (Kleywegt and Jones, 1997). Recently, the Jones group introduced EigenTHREADER, a novel fold recognition method which combines standard threading methods with their in-house MetaPSICOV contact prediction constraints method

(Buchan and Jones, 2017).

Phyre2 (Kelley et al., 2015) relies on HHsearch searches. Once templates have been identified, MODELLER is then used to predict the most probable model. Amino acid side chains are added to the final model using SCWRL4 (Krivov et al., 2009). In addition to the comparative modelling pipeline, Phyre2 also provides multiple-template and *ab initio* approaches to model the query. Recently, Phyre2 introduced PhyrePower which models queries with distant homology using contact threading i.e. pairwise alignment of eigendecomposed contact maps (Kelley, 2017).

Both SUPERFAMILY and Gene3D use HMMER3 (Eddy, 2011) to search their template libraries (based on SCOP and CATH respectively). Structural models are created by using the HMM alignment of the sequence to the best superfamily, and then resolved using MODELLER.

2.1.5 Objectives

As mentioned in Section 2.1.1, several approaches are used for identifying a close relative with known structure, to provide a template for comparative modelling. Where very close homologues are available ($\geq$30% sequence identity), it is possible to detect the closest template using the results returned by BLAST. However, when only remote homologues are available, it is best to scan against sequence profiles or HMMs constructed from closely related sets of homologues e.g. within a SCOP or CATH superfamily. The Orengo Group recently developed a sub-classification of CATH protein domain superfamilies that clusters relatives likely to have very similar structures and functions.

Functional families (FunFams) were introduced as a sub-classification of superfamilies inside the CATH-Gene3D, a resource which provides evolutionary classification of structures and sequences for known protein domains (Dawson et al., 2017; Lam et al., 2016). (See the Introduction chapter for a more detailed explaination of FunFams). When FunFams were used to select templates for building models of structurally uncharacterised relatives in 11 large, structurally and functionally diverse superfamilies, in the Structure Function Linkage Database (SFLD) (Akiva et al., 2014), the structural coverage of models was up to 5 times greater, for some superfamilies, compared to selecting targets using a 30% sequence identity cut-off. Furthermore, despite the fact that many remote homologues were used as templates, the final models were found to be of similar quality to those built using close sequence homologues ($\geq$30% sequence identity) as parents (Lee et al., 2010).

A recent, more accurate FunFam identification protocol (FunFHMMer, (Das et al., 2015)) uses similarities in sequence patterns, reflecting highly conserved positions and specificity-determining positions, to guide the sub-clustering and family detection. Functional purity of the new FunFams was demonstrated in a number of ways: validating against experimentally determined Enzyme Commission (Webb, 1992) and SFLD annotations, and also checking whether known functional sites coincide with highly conserved residues in the MSAs of FunFams (Das et al., 2015). Functional predictions based on FunFams were ranked amongst the top five methods for "Molecular Function" category and "Biological Process" category in the 2nd CAFA international function prediction experiment (Jiang et al., 2016).

In this chapter, we assess the quality of models built using a new modelling pipeline (FunMod) that uses the CATH-Gene3D FunFams to select parent templates. We first calculate the structural conservation of structural domains within FunFams. Since many model quality assessment methods exist, in order to assess the model quality, we first determine which of these methods is the best for our benchmark. After that, we assess which sequence alignment methods produce more good quality models. Then, we compare the quality of models built using FunFams to select templates with those built using HHsearch to identify templates. After that, we develop a modelling pipeline that combines different template strategies to model structurally uncharacterised proteins in human and fly genomes.

2.2 Materials and Methods

2.2.1 Calculating the structural conservation of structural domains within CATH-Gene3D FunFams and CATH superfamilies

To explore the conservation of domains within FunFams, we clustered all the CATH structural domains for each FunFam (which contains structural domains) into sequence identity 90% (S90) clusters. The structural representative for each S90 cluster was selected as the relative whose length is closest to the average length of the domains of the cluster. After that, we carried out all the pairwise structural comparisons between all the S90 representatives using SSAP (Taylor and Orengo, 1989).

We also compared the structural conservation of CATH domains across CATH superfamilies to compare against the conservation within FunFams. For each superfamily, we also compared representatives from sequence identity 35% (S35) clusters. The S35 structural representative was selected based on the average length of domains inside the S35 cluster and for having the best X-ray resolution of the structure. For each type of analysis (i.e. FunFam or superfamily), we carried out pairwise structural comparisons between all the representatives.

We calculated the mean of the normalised RMSD and SSAP score for the comparisons.

2.2.2 Assessing the model quality assessment methods, query-template alignment and template selection strategy

To assess the protocols for selecting templates, aligning templates and queries, ranking and assessing models, a query dataset was compiled from sequences of known structural domains in CATH (Sillitoe et al., 2015). The clustering program CD-HIT (Fu et al., 2012) was used to cluster CATH domain sequences into clusters of relatives showing 30% or more sequence identity (S30 clusters). A single representative was selected from each cluster to generate a non-redundant dataset.

To make sure the proteins selected have good structure, only proteins solved by X-ray crystallography with resolution equal to or better than 2Å were used. At this resolution, proteins tend to have fewer errors with only a few incorrect rotamers (Huang, 2007).

This dataset was separated into two subsets based on sequence homology. The close homologue subset consists of query targets that have sequence relatives with more than or equal to 30% global sequence identity. By contrast, the remote homologue subset comprises query targets that have sequence relatives with less than 30% global sequence identity. This gave 8,633 close homologue query targets and 602 remote homologue query targets.

2.2.2.1 Template selection methods

The default parameters suggested by the authors of the respective methods were used for all the methods.

HHsearch Query sequences in the benchmark set were scanned against the CATH HHsearch HMM library using HHsearch. The library is composed of HMM built for every single domain in the CATH version 4.0. The HMM library was built according to the protocol provided in the HHseach guidebook: For each CATH target sequence, we performed two iterations of HHsearch searches against the UniProt20 database (provided by HHsearch, clustered at 20% sequence identity), with an E-value threshold of 0.001.

We selected the structural template with the highest probability to be a true positive provided by the program, which considers both the sequence and secondary structure alignment for each target sequence. The structural templates selected must exceed the default probability score of 20 and E-value below 0.001. The original structures of the query sequences were excluded as templates.

FunFam Query sequences in the benchmark set were scanned against the HMMs for each CATH FunFam using HMMER3. The FunFam of the best-matched HMM was selected. If the matched FunFam did not contain any structural templates, the next closest FunFam was checked to determine whether there was a structural template. However, the query sequence had to match the FunFam with an E-value threshold of <0.001.

The original structures of the query sequences were excluded as templates. The best templates were selected based on the average BLOSUM62 sequence similarity with the query targets. Templates with X-ray resolution below 5Å were prioritised.

2.2.2.2 Comparing different model quality assessment methods

A variety of model quality assessment methods were tested. These methods were used to select the best models from ten models built by MODELLER. nDOPE (i.e. normalised DOPE) (Shen and Sali, 2006), GOAP (Zhou and Skolnick, 2011), and BACH (Sarti et al., 2013) are statistical potential based single model quality assessment methods. Whereas, GA341 combines statistical potentials with information on target-template sequence identity, and structural compactness (Melo et al., 2002). ProQ2 (Ray et al., 2012) is a machine learning method that is based on evolutionary information, a MSA and structural features. ModFOLDclust2 (McGuffin and Roche, 2010) is a clustering based model quality assessment method.

To assess the performance of the model assessment methods, the best model selected by the different methods was superposed on the original protein structure, and the quality of the model was given by the structural similarity score.

2.2.2.3 Comparing different query-template alignments methods

MAFFT alignment We employed MAFFT MSA method (Katoh and Standley, 2013) to align the query and template identified by the different template searching algorithms.

For HHsearch, we realigned the sequences of the best HHsearch match using MAFFT (L-INS-i mode). Next, the query sequence was added to the alignment using MAFFT (using the mafft –add option). Subsequently, for both BLAST and HHsearch, the alignment of query and template was submitted to MODELLER.

For FunFam method, query sequences were added to the pre-built FunFam alignment of the matched FunFams using MAFFT and the alignment of query and template was submitted to MODELLER. The original FunFam alignment was produced using MAFFT (L-INS-i mode).

HHsearch alignment For HHsearch, we obtained the query-template sequence alignment provided by the HHsearch program. For the FunFam method, after identifying the FunFam to which the query belongs and selecting the best structural templates, we obtained an HHsearch HMM profile for the structural template to generate an alignment. Then, we aligned the query sequence with the template HMM profile using the HHsearch alignment methods and extracted the query-template alignment.

2.2.2.4 Comparative modelling

Comparative modelling software MODELLER version 9.15 was used to predict ten models for each query target for each template selection method. The best model was selected using the best model quality assessment score obtained in Section 2.3.2.

2.2.2.5 Assessing the performance of the prediction protocols and ranking the models

In order to assess the target selection, sequence alignment and model ranking protocols, TMscore, a sequence dependent structural superposition program (Zhang and Skolnick, 2005), was used. TMScore calculates a TM-score (Zhang and Skolnick, 2004) in the range of 0 to 1. Protein pairs with TM-score $>$0.5 are generally in the same CATH/SCOP fold group (Xu and Zhang, 2010).

The TMScore program was also used to calculate the GDT-HA score. The GDT-HA score takes into account the number of Cα pairs within a distance 0.5Å, 1Å, 2Å, and 4Å after superposition of the two structures (Read and Chavali, 2007; Zemla, 2003; Zemla et al., 1999). GDT-HA is the current official structural comparison score used in The Critical Assessment of Protein Structure Prediction (CASP) competition (Kryshtafovych et al., 2015).

2.2.3 Modelling structurally uncharacterised sequences in human and fly

Once a robust modelling pipeline had been established by the benchmarking, 3D models were generated for human and fly domain sequences classified in the CATH-Gene3D resource (version 12 (Lees et al., 2014)). After removing all sequences that already had a structure in the PDB, there were 97,326 human (*Homo sapiens*) sequences and 36,761 fly (*Drosophila melanogaster*) sequences inclusive of isoforms.

Different template searching methods were combined to model all sequences:

- FunFams
- HHsearch (using the CATH 4.0 HMM library, which includes HMMs for all CATH domains)
- HHsearch (using the SCOP 1.75 HMM library clustered at 70% sequence identity)
- HHsearch (using the SCOe 2.05 HMM library clustered at 90% sequence identity)
- HHsearch (using the September 2014 release of PDB HMM library clustered at 70% sequence identity)

After the models had been built, we used the best model quality assessment score obtained in Section 2.3.2 to assess the quality of the models.

2.3 Results and Discussion

We first explored the structural conservation of domains within FunFams and compared this with the structural conservation of domains across the superfamily.

2.3.1 Structural conservation of structural domains within FunFams and across superfamilies

We calculated the structural conservation for a total of 1,307 CATH superfamilies and 5,675 FunFams. Figure 2.2 contains two histograms that demonstrate the nRMSD and SSAP score differences between structural domains within FunFams and across Superfamilies. The majority of FunFam structural domains have a nRMSD value between 0 to 5Å and a SSAP score between 80 to 100. By contrast, for pairs of superfamily domains, the nRMSD values have a wider spread from 0 to 10Å and SSAP score differences between 70 and 90. Both the SSAP and nRMSD differences between the groups were statistically significant (p-value $<$2.2e-16, Mann-Whitney U test). This demonstrates that the structural domains within FunFams are more structurally conserved than within Superfamilies.

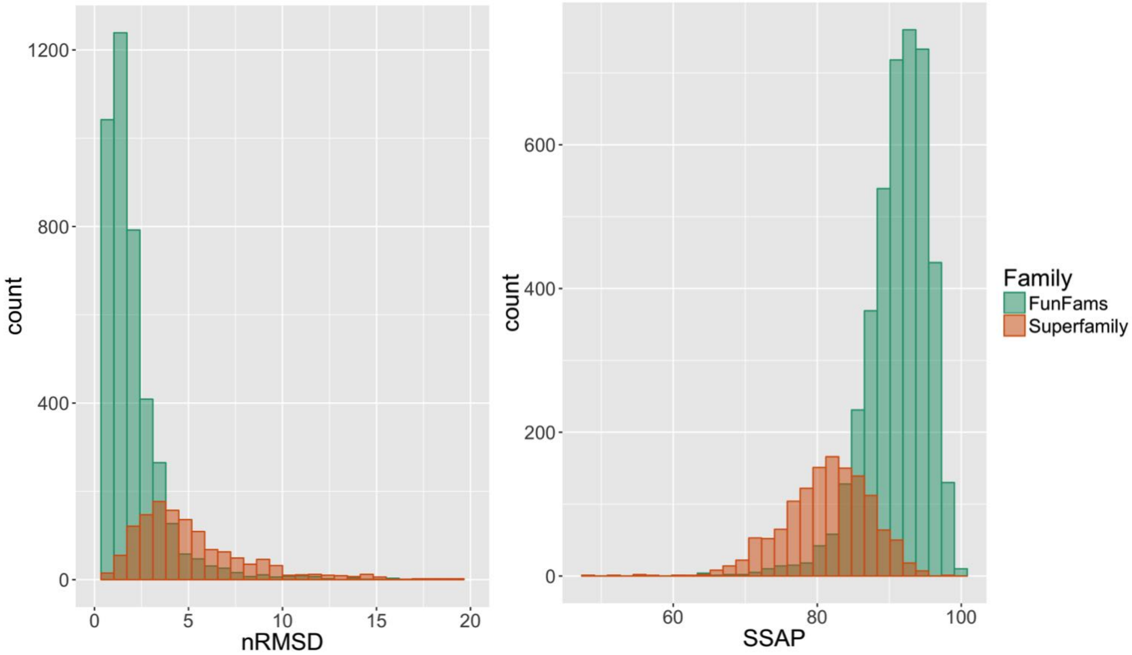


Figure 2.2: Structural conservation of domains within FunFams and Superfamilies.

2.3.2 Choosing the best model assessment method

We first tested different model assessment methods to determine/establish their performance. This quality of the models was assessed directly by superposing them on the native structures for the target sequences. For each target, top ranked models (out of the 10 models built) were selected by different model assessment methods.

Figure 2.3 illustrates the qualities of top ranked models selected by different model assessment methods. The top ranked models have similar quality regardless of which structure comparison methods were used to assess performance.

Since all the methods gave comparable performance in ranking models, MODELLER's built in method nDOPE was used to select the best model in the studies reported below, as this was the most convenient option.

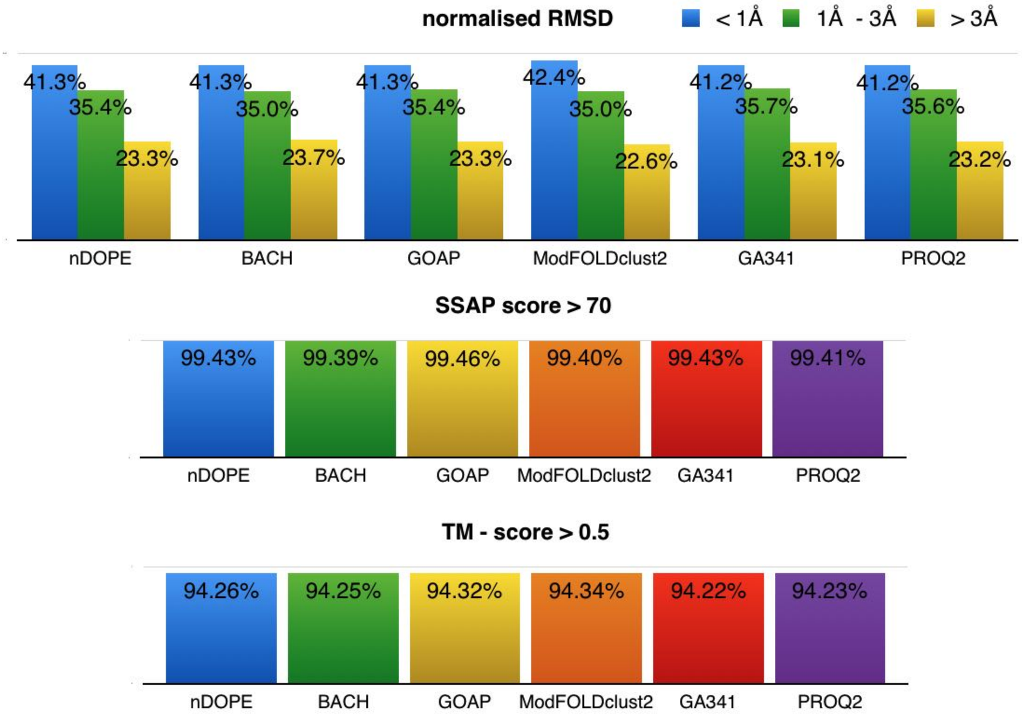


Figure 2.3: Comparison between different model quality assessment methods. The percentage of top ranked models (selected by a particular method) with a certain degree of structural similarity with the native structure are shown.

2.3.3 Choosing the best query-template alignment method

There are 3 major steps to the modelling process: (1) Find the template (2) Align the query and template (3) Build the model. We first tested the different strategies for aligning the query and template. These strategies were assessed by building the models and assessing the model quality.

We used the HHsearch template selection method in this analysis. We examined the performance of the two different alignments methods: HHsearch default query-template alignment and MAFFT query-template alignment. Alignments were compared using the same structural templates for the same query sequences. HHsearch aligns the query sequence to the sequences in the HMM model to which the query sequence has the best match. For MAFFT, the query sequence is matched to a HHsearch HMM and then align the sequences from the HMM model aligned with the query. Figure 2.4 demonstrates the percentages of good quality models built for both close homologs (sequence identity ≥30%) and remote homologs (sequence identity <30%) for the alignment methods. Overall, the HHsearch alignment method produced a higher proportion of good quality models than the MAFFT alignment method.

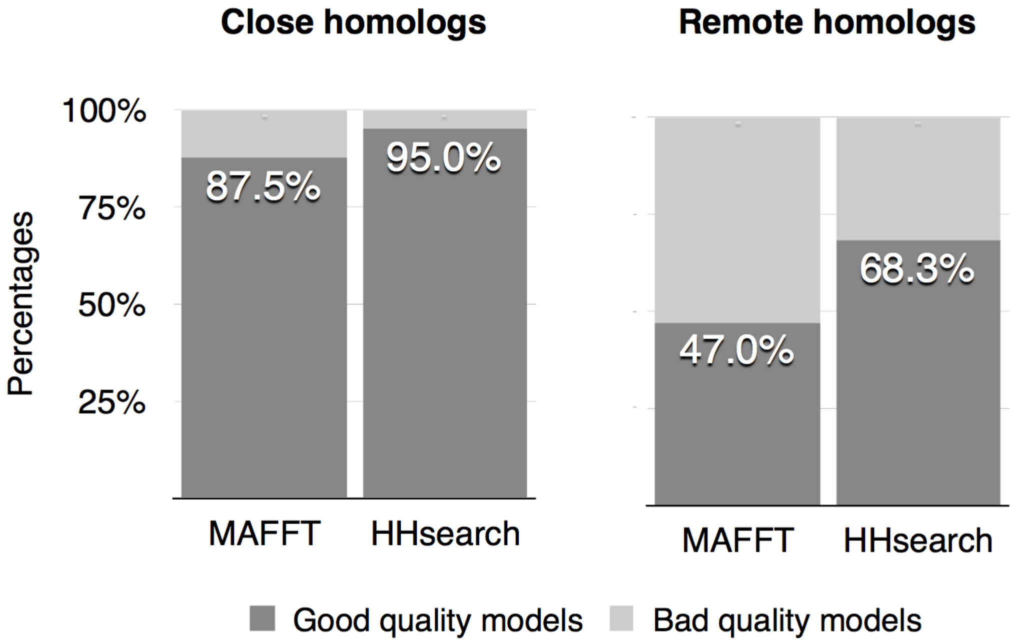
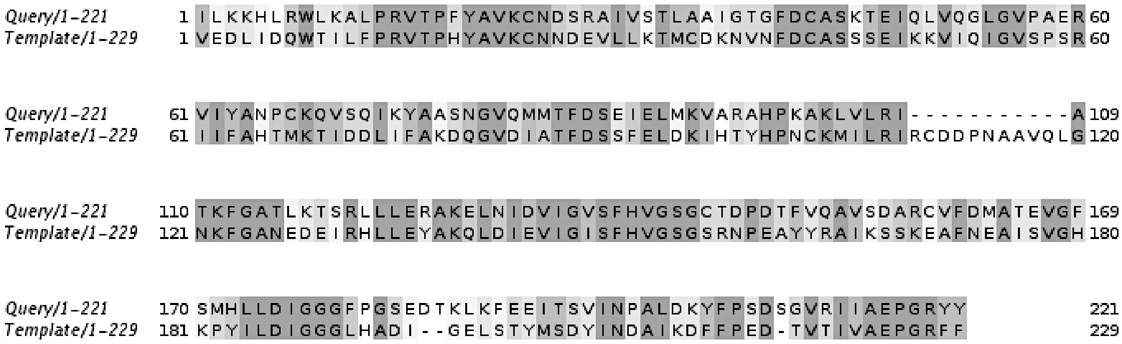


Figure 2.4: Comparison between different sequence alignment methods. Good models are defined by models with TM-score >0.5 when compared to the native structure.

We examined one extreme example where the two query-template alignments produced are very different (See Figure 2.5, Figure 2.6 and Figure 2.7). The sequence identity between the template 2nv9F02 and the query 7odcA02 was 39%. The HHsearch query-template alignment is good. However for the MAFFT alignment, most of the residues are not aligned properly and lots of gaps are added. Problems in the alignment are reflected in the models produced. HHsearch employs an alignment algorithm that maximizing the number of correctly aligned pairs of residues (i.e. the posterior probability of match state to be aligned). HHsearch also has an default filtering step that removes sequences that are too distant. The MAFFT strategy does not include a filtering step and so MAFFT will be aligning many very distant relatives.



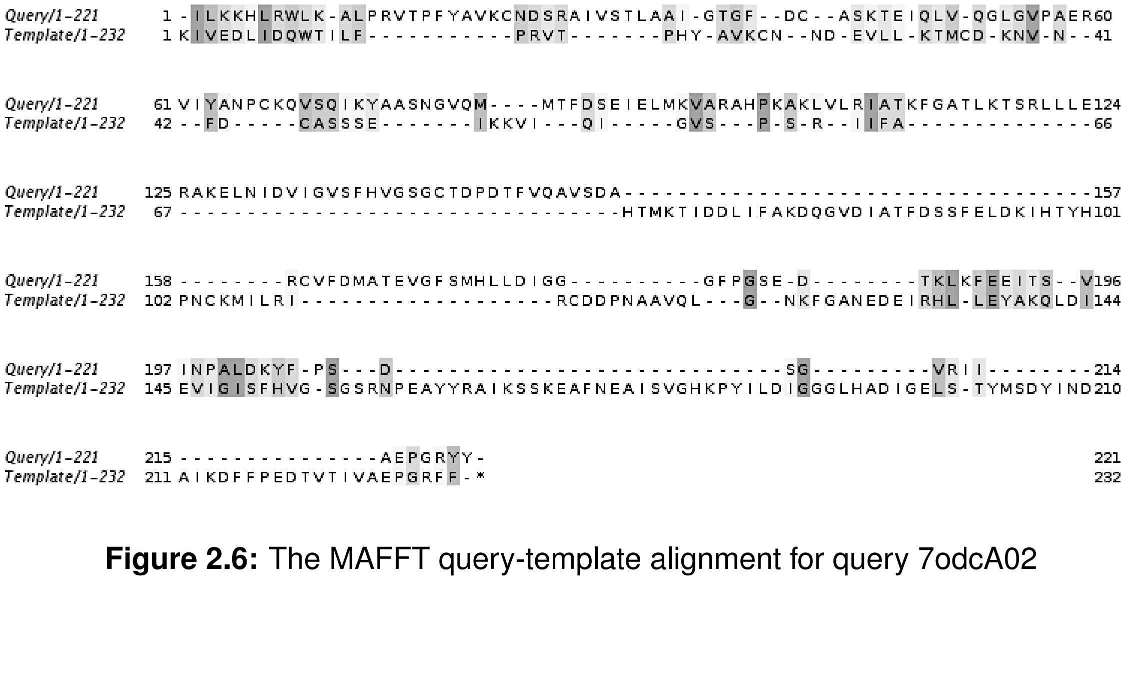
```
Query/1-221        1 ILKKHLRWLKALPRVTPFYAVKCNDSRAIVSTLAAIGTGFDCASKTEIQLVQGLGVPAER 60
Template/1-229     1 VEDLIDQWTILFPRVTPHYAVKCNNDEVLLKTMCDKNVNFDCASSSEIKKVIQIGVSPSR 60

Query/1-221       61 VIYANPCKQVSQIKYAASNGVQMMTFDSEIELMKVARAHPKAKLVLRI-----------A 109
Template/1-229    61 IIFAHTMKTIDDLIFAKDQGVDIATFDSSFELDKIHTYHPNCKMILRIRCDDPNAAVQLG 120

Query/1-221      110 TKFGATLKTSRLLLERAKELNIDVIGVSFHVGSGCTDPDTFVQAVSDARCVFDMATEVGF 169
Template/1-229   121 NKFGANEDEIRHLLEYAKQLDIEVIGISFHVGSGSRNPEAYYRAIKSSKEAFNEAISVGH 180

Query/1-221      170 SMHLLDIGGGFPGSEDTKLKFEEITSVINPALDKYFPSDSGVRIIAEPGRYY         221
Template/1-229   181 KPYILDIGGGLHADI--GELSTYMSDYINDAIKDFFPED-TVTIVAEPGRFF         229
```

Figure 2.5: The HHsearch query-template alignment for query 7odcA02



```
Query/1-221        1 -ILKKHLRWLK-ALPRVTPFYAVKCNDSRAIVSTLAAI-GTGF--DC--ASKTEIQLV-QGLGVPAER60
Template/1-232     1 KIVEDLIDQWTILF-----------PRVT-------PHY-AVKCN--ND-EVLL-KTMCD-KNV-N--41

Query/1-221       61 VIYANPCKQVSQIKYAASNGVQM----MTFDSEIELMKVARAHPKAKLVLRIATKFGATLKTSRLLLE124
Template/1-232    42 --FD-----CASSSE-------IKKVI---QI-----GVS---P-S-R--IIFA--------------66

Query/1-221      125 RAKELNIDVIGVSFHVGSGCTDPDTFVQAVSDA-----------------------------------157
Template/1-232    67 ---------------------------------HTMKTIDDLIFAKDQGVDIATFDSSFELDKIHTYH101

Query/1-221      158 --------RCVFDMATEVGFSMHLLDIGG----------GFPGSE-D--------TKLKFEEITS--V196
Template/1-232   102 PNCKMILRI-------------------RCDDPNAAVQL---G--NKFGANEDEIRHL-LEYAKQLDI144

Query/1-221      197 INPALDKYF-PS---D-----------------------------SG---------VRII--------214
Template/1-232   145 EVIGISFHVG-SGSRNPEAYYRAIKSSKEAFNEAISVGHKPYILDIGGGLHADIGELS-TYMSDYIND210

Query/1-221      215 --------------AEPGRYY-                                              221
Template/1-232   211 AIKDFFPEDTVTIVAEPGRFF-*                                             232
```

Figure 2.6: The MAFFT query-template alignment for query 7odcA02

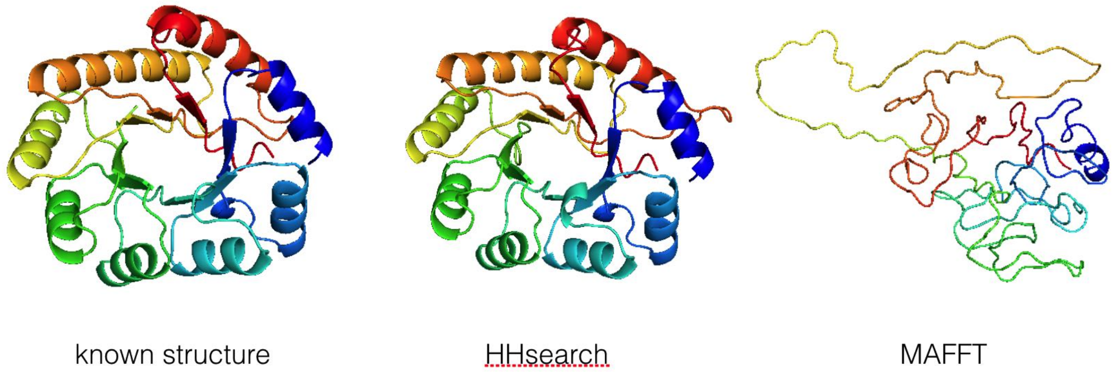


Figure 2.7: Known structure and structural models built using HHsearch and MAFFT alignments for 7odcA02

Since this analysis demonstrated that the HHsearch query-template alignment performed better than MAFFT in the modelling pipeline, the HHsearch method was used in the following analyses of template selection protocols.

2.3.4 Assessing the performance of FunFam and HHsearch template selection methods

In the previous section, we demonstrated that the HHsearch alignment strategy works better for comparative modelling. Here we assessed the different template selection strategies (i.e. HHsearch and FunFams). Again, this was done by building the models and assessing the model quality.

Figure 2.8 demonstrates the proportion of good quality models built by FunFams and HHsearch. FunFams gave a high proportion of good quality models than HHsearch for both close (sequence identity $\geq$30%) and remote homologs (sequence identity $<$30%). The difference was statistically significant for models built for remote homologs (p-value $<$1E-19, Mann-Whitney U test).

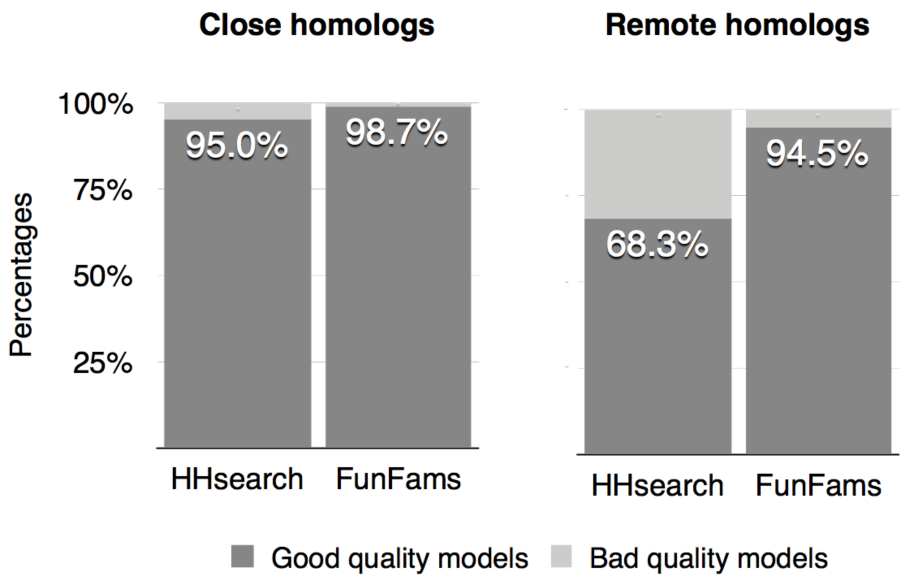


Figure 2.8: Proportion of good quality models built by FunFam and HHsearch.

2.3.4.1 Close homologues with sequence identity ≥50%

Figure 2.9 shows that both template selection strategies gave similar numbers of good quality models. This result is not surprising as the FunFam and HHsearch protocols either selected the same template or another closely related structural template. At this level of sequence homology, sequences tend to share high structural similarity, so choosing an alternative close homologue as a template is unlikely to affect the quality of the models built.

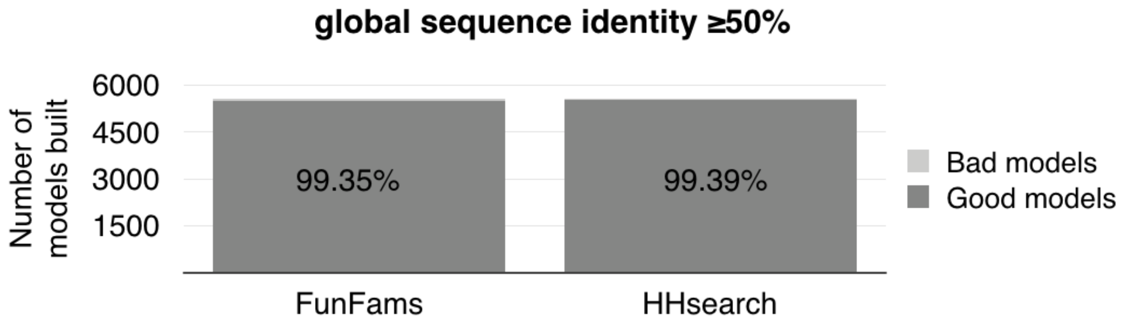


Figure 2.9: Number of models built from templates selected by the FunFam and HHsearch protocols for homologues with sequence identity ≥50%. Good models are defined by models with TM-score >0.5 when compared to the native structure.

2.3.4.2 Close homologues with sequence identity 30%-50%

Figure 2.10 compares the performance of the FunFam protocol versus the HHsearch protocol, respectively, for homologues in the sequence identity range 30%-50%. When comparing target selection and alignment protocols, for each query target, models were assigned to one of the following three categories: (1) models that were produced from the same template (2) models that were generated by different templates (3) extra models that could only be built by a particular method.

Overall, the performance of the FunFam protocol is comparable to the HHsearch protocol. The HHsearch protocol gave 4 more good quality models than the FunFam protocol. Both of the methods managed to identify some query targets, which the other protocol failed to identify. FunFams built an extra 46 models and HHsearch built an extra 593 models. 84.8% of these extra FunFams models are of good quality, and 44.3% of HHsearch models are good. FunFams identifies fewer targets because the protocol only allows models to be built if there is a highly confident match.

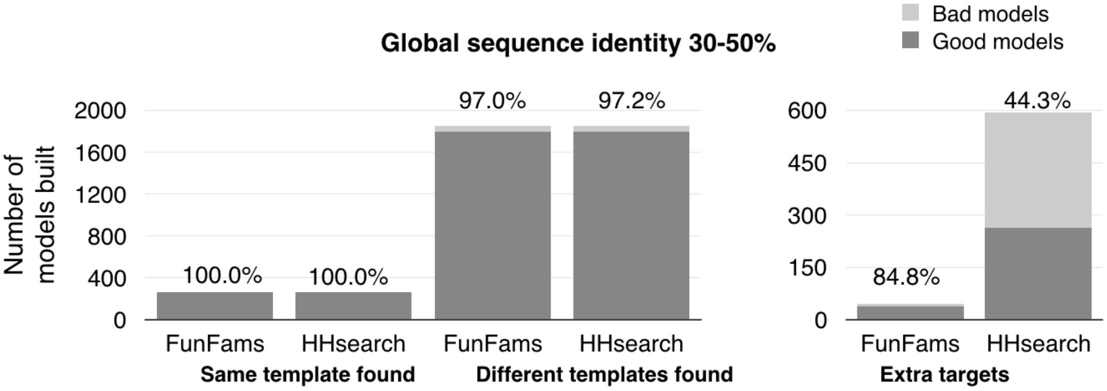


Figure 2.10: Number of models built by the FunFam and HHsearch protocols for homologues with sequence identity 30%-50%. Good models are defined by models with TM-score >0.5 when compared to the native structure.

2.3.4.3 Remote homologues with sequence identity <30% sequence identity

Figure 2.11 demonstrates the quality of models built by the FunFam protocol and the HHsearch protocol for the remote homologues in the query dataset. FunFam gave slightly more good models than HHsearch protocol for the common models.

HHsearch managed to identify 220 templates not selected by FunFams. However, 56% of the models built are low quality models. We observed a similar phenomenon with close homologues with sequence identity 30%-50%. The HHsearch protocol tends to model more targets than the FunFam protocol, but about half of the models built are of low quality. The FunFam protocol gave fewer models but a higher proportion of good quality models.

Using this benchmark dataset, it appears that FunFams do not identify any additional templates compared to the HHsearch strategy.

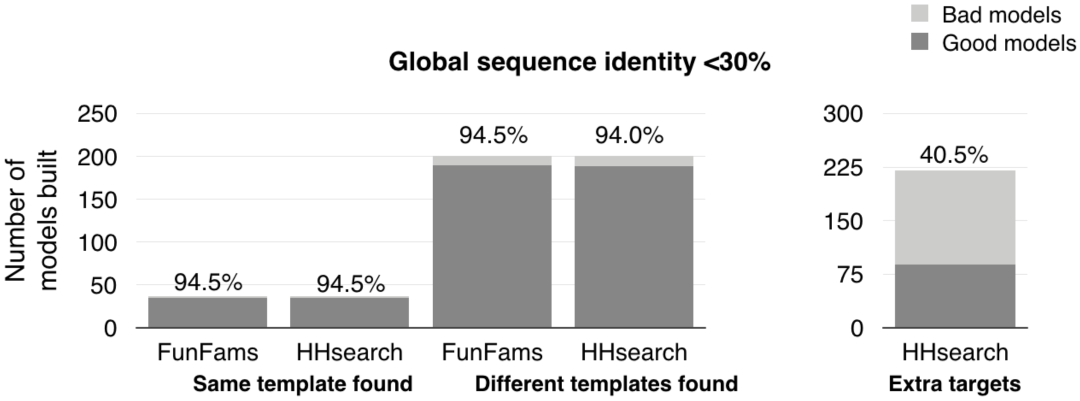


Figure 2.11: Number of models built by the FunFam and HHsearch protocols for remote homologues. Good models are defined by models with TM-score $>$0.5 when compared to the native structure.

Figure 2.12 demonstrates the distribution of model quality for models built for queries using different templates. Overall, FunFams and HHsearch models are comparable. However, there are slight differences depending on the method used to assess model quality. FunFam models score slightly better with GDT-HA, whereas HHsearch models are slightly better assessed using the TM-score. TM-score is a global structural comparison score that accounts for all the residues of the modelled proteins, GDT-HA uses distance cut-offs and focuses on fractions of the structures that are correctly modelled.

Therefore, HHsearch models have better global similarity with the native structure and FunFams models tend to have a higher local agreement with the native structure. Having a local similarity is important when we are modelling enzymes or protein complexes, where a better representation of functional sites is crucial.

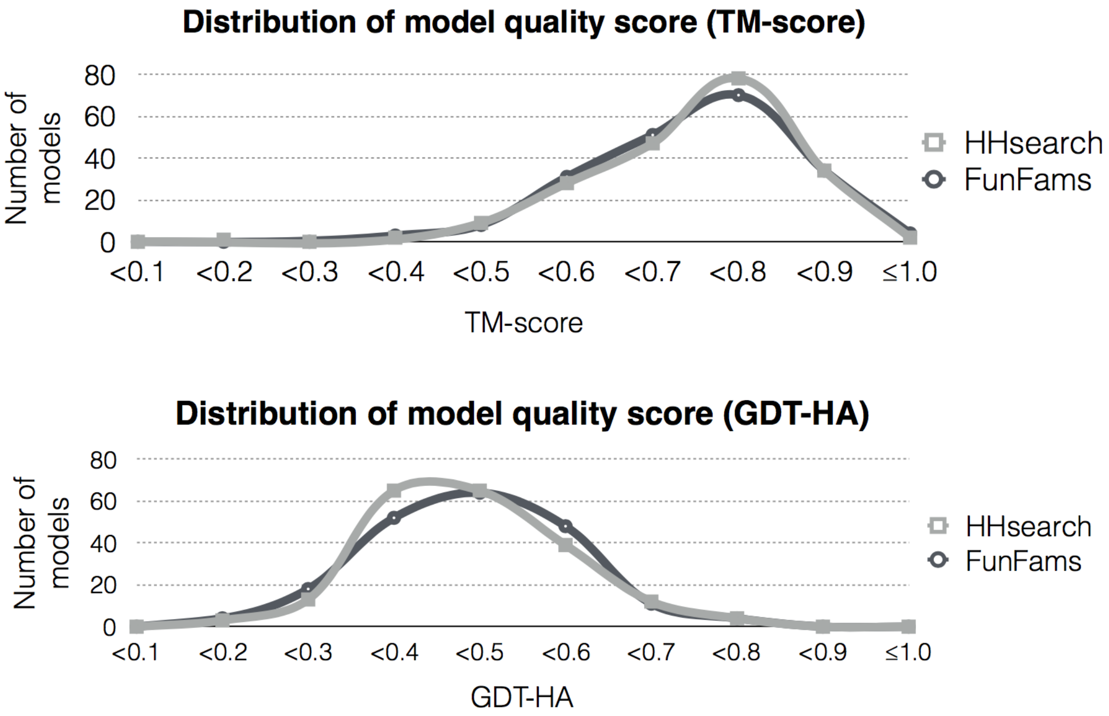


Figure 2.12: Distribution of model quality scores of common remote models (using different templates) built by the FunFam and the HHsearch protocols. Similar structures gave higher GDT-HA/TM-scores.

2.3.4.4 Which protocol selects a higher proportion of good templates than the other protocol

We carried out an analysis to determine how often the FunFam protocol or the HHsearch protocol selected a better template (compared to the other). To identify which protocol selects the best template, we performed a structural comparison between the structural templates, against the query structure. We compared 5,977 close and 146 remote cases where the protocols selected different structural templates and found ∼80% of the chosen FunFam and HHseach templates had nRMSD score below 3Å when compared to the query structure.

We subtracted the nRMSD value of FunFam structural comparison score by the nRMSD value of HHsearch structural comparison score to determine which protocol selects better template. Table 2.1 demonstrates the nRMSD difference of the templates selected. We observed that in 67% of the cases FunFam and HHsearch select structurally similar templates. There is a slight tendency for the FunFam protocol to select better structural templates than the HHsearch protocol. This is statistically significant (p-value $<$ 2E-12, Wilcoxon signed ranked test).

Table 2.1: How often do the two protocols select a better template?

ΔnRMSD$\geq$1Å (HHsearch selects better templates)	13%
1Å$<\Delta$nRMSD$<$1Å (Similar templates)	67%
ΔnRMSD$\leq$-1Å (FunFam selects better templates)	20%

2.3.5 Modelling human and fly genomes

In the previous section, we demonstrated the ability of the FunFam protocol to build a higher proportion of good quality models than the HHsearch protocol. In this section, we report a modelling pipeline that utilised both the FunFam protocol and the HHsearch protocol to model structurally uncharacterised human and fly sequences found in the CATH-Gene3D resource. Different HHsearch libraries based on CATH, SCOP, SCOPe and PDB seed structures were used. Figure 2.13 illustrates the FunMod modelling pipeline.

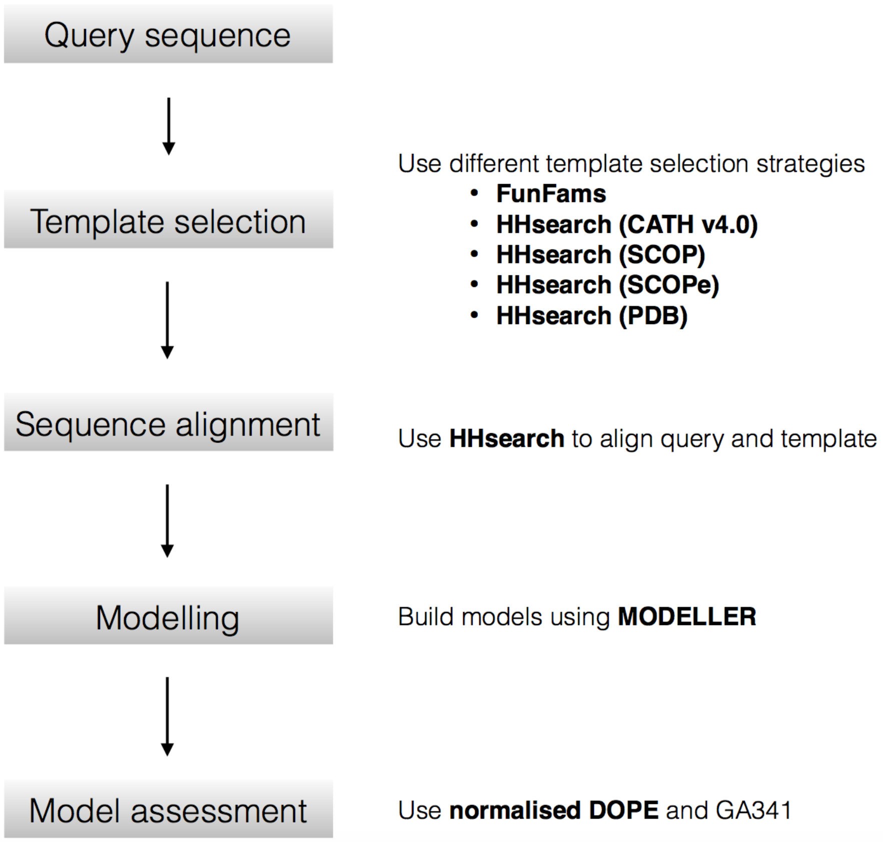


Figure 2.13: FunMod modelling pipeline

2.3.5.1 Analysis of model quality for human and fly genes using FunFams and HHsearch (CATH)

Our previous analysis in Section 2.3.2 showed that all the model quality assessment methods perform equally well. In this study, we employed nDOPE and GA341 to assess the quality of the models. Both of these scores are widely applied and are used by the comparative modelling resource ModBase to assess the model quality (Eswar et al., 2008; Pieper et al., 2014). Good quality models are models that score an nDOPE score below 0 and GA-341 score above 0.7 with a false positive rate of 3.25% (Melo and Sali, 2007).

Note, the HHsearch(CATH) strategy is the HHsearch protocol we have discussed in Section 2.3.4. Both the FunFam and HHsearch uses the same HHsearch alignment strategy. Figure 2.14 demonstrates the number of good quality human and fly models built by both HHsearch(CATH) and FunFams protocols. Both protocols model a significant proportion of the structurally uncharacterised genes in both model organisms. For approximately 21% of the proteins only one of the protocol could produce a good model. As expected, the HHsearch (CATH) protocol gave more models than the FunFams protocol. Yet, FunFams did give a considerable amount of extra models.

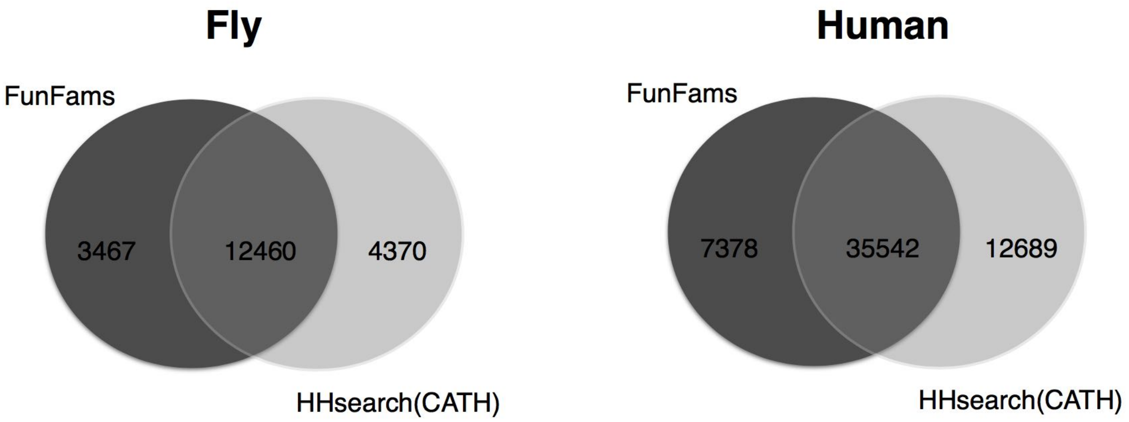


Figure 2.14: Number of good quality FunFams and HHsearch (CATH) human and fly models

For all the common models built by the two protocols, we investigated the predicted quality of the models based on nDOPE. Overall, we can see that FunFams models gave a higher nDOPE compared to HHsearch models. The differences were significant (<2.2e-16, Wilcoxon signed-rank test). This demonstrates that generally FunFam models are more likely to be good quality than the HHsearch(CATH) models (See Figure 2.15).

The result differs slightly from our previous analysis (see Section 2.3.4) due to the different query dataset used. The benchmark dataset in the previous section is a selection of representative CATH structures which capture the structural diversity in the CATH resource. The data set in this analysis is possibly more representative of protein space.

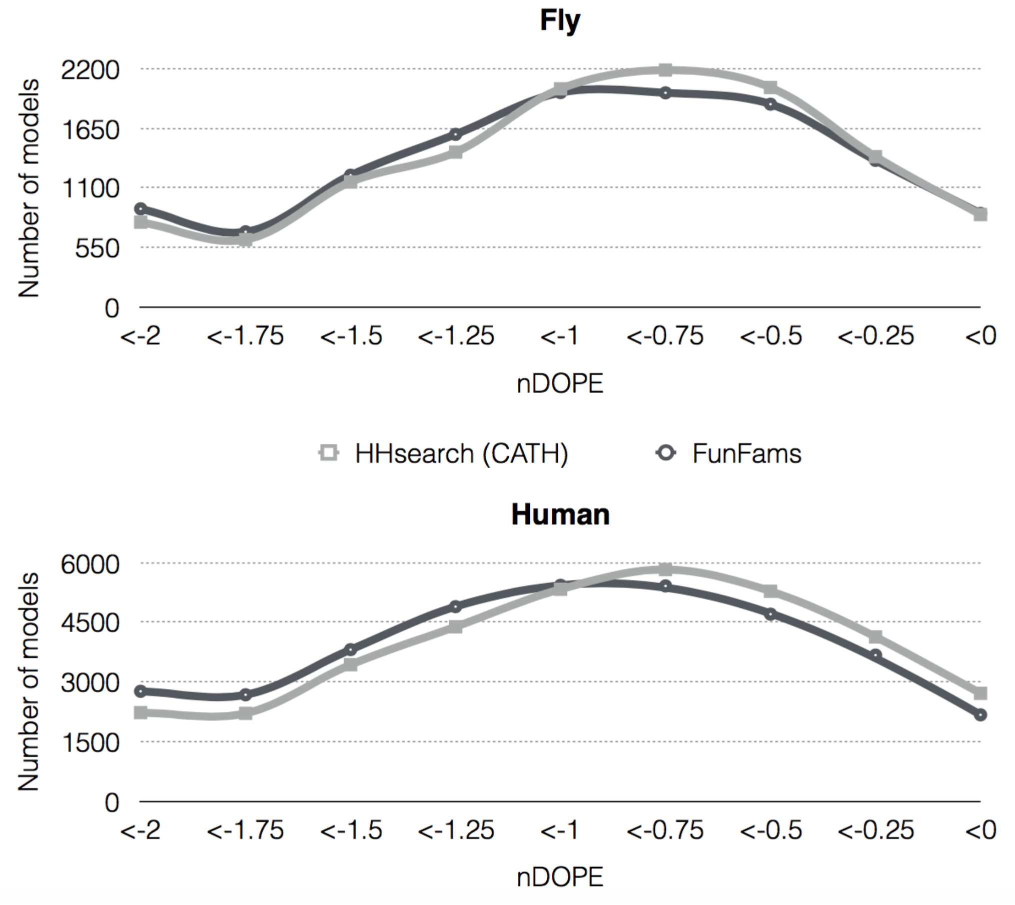


Figure 2.15: Distribution of model quality scores for common FunFams and HHsearch (CATH) human and fly models. Good quality models give lower nDOPE score.

In summary, this analysis demonstrates that the FunFam protocol gave slightly more good quality models than the HHsearch (CATH) protocol. This means that the FunFam protocol has a lower false discovery rate compared to the HHsearch (CATH) protocol.

2.3.5.2 Analysis involved FunFams and all the HHsearch libraries

Figure 2.16 demonstrates the number of good models built by the different strategies. One thing to mention, only 70% of the CATH domains are included in FunFams. Therefore, the HHsearch (CATH) libraries covers a lot of CATH domains than FunFams.

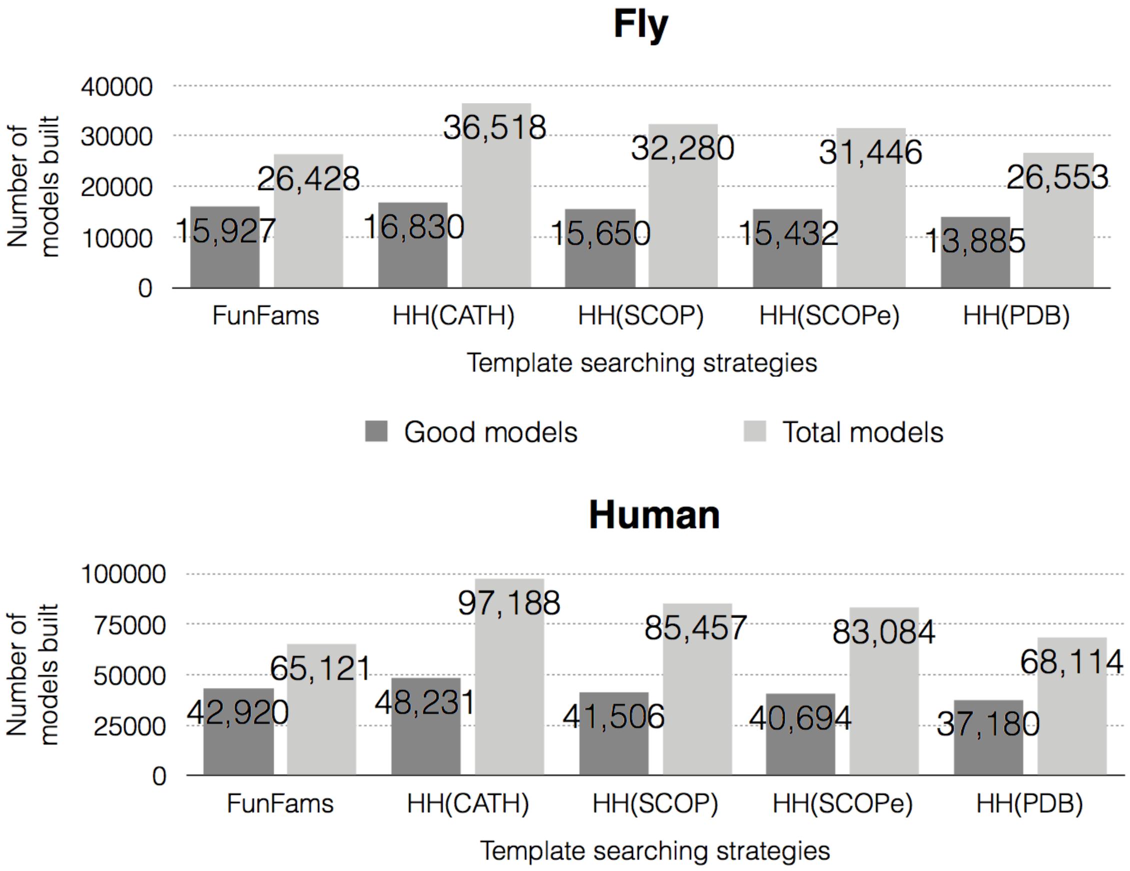


Figure 2.16: Number of models built by the different template searching strategies

The Venn diagram in Figure 2.17 summarises the common and extra models built by the different strategies. All strategies produce a significant number of models and there are many query sequences for which only one strategy can build a model. Therefore it is valuable to use multiple search strategies and template libraries to model the whole genome.

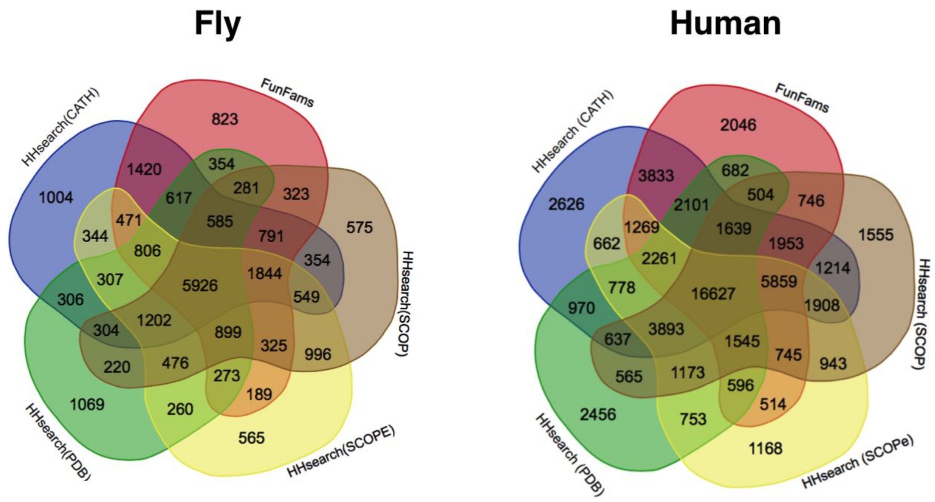


Figure 2.17: Overlap of good models built by different strategies

We selected the best models built for every Gene3D sequences based on the model quality. By applying this strategy, we extended the structural coverage from 10% known structure to nearly 70% for both fly and human sequences (Figure 2.18).

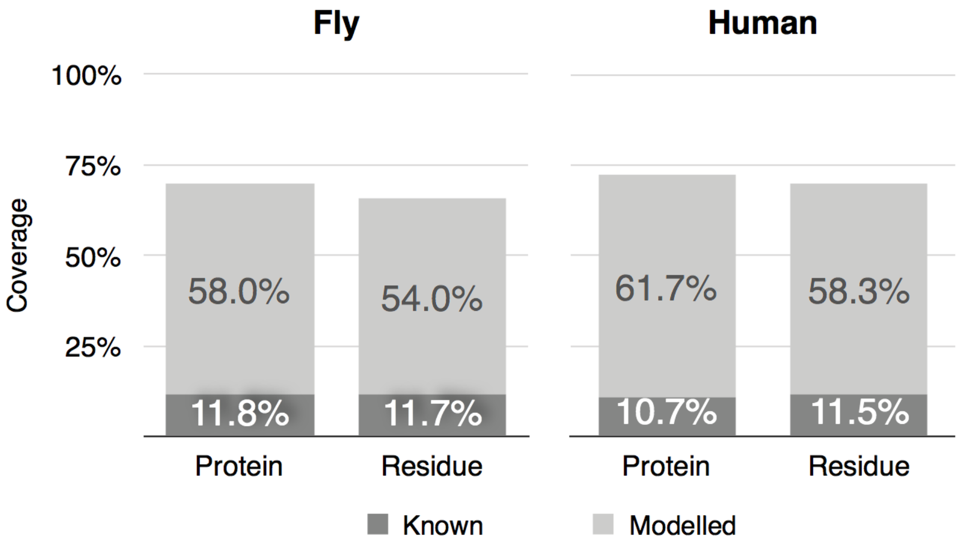


Figure 2.18: Structural coverage of human and fly CATH-Gene3D sequences

2.4 Conclusions

Comparative modelling performance usually depends on three steps: (1) template selection (2) sequence alignment between template and query (3) comparative modelling. This chapter mainly focused on the template selection and sequence alignment process. These two steps are the most critical since if a wrong template is selected or if the right template is badly aligned, it is unlikely to yield a good model.

Although many new model quality assessment protocols (MQAPs) have been developed recently, we found that all performed similarly in selecting the best model from sets of models generated by MODELLER. We, therefore, used MODELLER's built in program, DOPE, to rank models.

We explored whether the FunFam template selection protocol identified better template structures than the HHsearch protocol. The FunFam protocol first assigns a query sequence to a FunFam and then selects the best template from the FunFam based on the sequence identity and X-ray resolution. For the HHsearch protocol, we used HHsearch to scan against a library of HMMs (built from all CATH domain structures (version 4.0)) and select the best template based on the program's built-in statistical measures. We employed the same alignment strategy for both protocols. The FunFam protocol gave a higher percentage of good models compared with HHsearch for both close homologues [95.0% (HHsearch) versus 98.7% (FunFams)] and remote homologues [68.3% (HHsearch) versus 94.5% (FunFams), p-value $<$1E-19; Mann-Whitney U test]. This suggests that it is helpful to subclassify homologues according to likely structural and functional similarity for the template selection process.

In addition, we developed a modelling pipeline that utilises the FunFam template search protocol and other template searching methods (i.e. HHsearch protocol utilising different template libraries). By combining these methods, we managed to model up to 70% of human and fly genomes with at least one structural domain model. We demonstrated that models built on templates identified by the FunFam protocol are more likely to be good quality compared to HHsearch (CATH) models (based on the model quality assessment score nDOPE). We also show the importance of using different template libraries to extend the structural coverage of fly and human genomes.

2.4.1 Future directions

Another major application of our modelling pipeline will be for protein complex modelling which will be discussed in Chapter 3. Protein modelling of structurally uncharacterised protein monomers can increase the number of protein complexes which can be built.

In the future, it would be valuable to test whether we can improve the sequence alignment between template and query using newer sequence alignment methods such as MRF-Align (which is based on Markov Random Forest) (Ma et al., 2014) that have been proven to produce better sequence alignments.

We have observed recently that some of the FunFams are polluted with sequences that are annotated with diverse GO terms, suggesting that they should not to be clustered in the same FunFam. Ongoing research work in the Orengo group is involved with sub-classifying these polluted FunFams into functionally pure clusters.

We can also incorporate secondary structure information into the FunFam HMMs. Secondary structure prediction programs such as PSI-PRED (Jones, 1999) can be used to predict the overall secondary structure profile for all the FunFam HMMs. This may increase the power of the template search as HMM with secondary structure information is stronger than a purely sequence-based HMM. The effect will be more profound if we can add secondary structure information in the early stages of FunFam building.

Also improvement in the loop and side chain modelling can be explored as well using methods such as using a better energy scoring function or a better side chain or rotamer library (Della Corte et al., 2016; Feig, 2016; Kim and Kihara, 2016; Lee et al., 2016; Park et al., 2016).

2.4.2 Uses of structural modelling in experimental studies

Below, we highlight a few selected examples of recent developments in techniques that exploit comparative models to improve the structural determination or structural coverage of large-scale macromolecular assemblies.

2.4.2.1 Facilitation of cryo-EM density map fitting with homology models

New developments in cryo-electron microscopy (cryo-EM) have meant that this approach is increasingly used for the protein structure determination of large macromolecular complexes and assemblies. One major problem with cryo-EM is the low resolution of the density maps that are produced. To help with the interpretation of these density maps, they are usually fitted onto experimentally solved structures. However, owing to the low number of solved structures, it can sometimes be hard to find a suitable template. In 2005, the Topf group demonstrated that it is feasible to use comparative models for the fitting process. They subsequently developed a web server named CHOYCE (Rawi et al., 2010) which performs homology modelling (MODELLER) and fitting into cryo-EM maps. The server allows the user to select the most accurate models (based on the DOPE score).

For those adventurous users who prefer to perform the modelling manually, Allen and Stokes exemplified the steps involved from building the structural models to the fitting of models to the density map using an integral membrane protein, CopA. In addition to this, they also illustrated how to dock additional components into the models using a computational approach (Allen and Stokes, 2013).

Gorgon (Baker et al., 2016) can model not only a protein structure but entire macromolecular assemblies. For example, the C backbone model for every protein component in the ribosome (from an 4.5Å resolution cryo-EM map) was automatically built in less than a day. Gorgon uses *ab initio* modelling, feature extraction and rigid-body and flexible fitting for model building. It also includes the use of statistical measures to evaluate the fit of an atomic model to the cryo-EM density map.

2.4.2.2 Integrative structural biology

Integrative structural biology is a new field which tries to determine the three-dimensional structures of proteins by using the ensembles produced by experimental methods and computational approaches (Ward et al., 2013). This is especially useful for proteins that are not crystallizable, are insoluble, are too large or too small or are conformationally heterogeneous (Sali et al., 2015).

Shi and coworkers used a refined integrative method that combines information generated from electron microscopy, X-ray crystallography and comparative structure modelling to provide a clear structural view of the Nup84 nucleoporin complex. This complex is a stable heteroheptameric (seven nucleoporins) protein complex of 600 kDa from budding yeast (Shi et al., 2014).

Another interesting example is the structure of human prolactin receptor solved by Bugge and coworkers in 2016. This was the first ever full view of a class I cytokine receptor. Class I cytokine receptors are generally considered to be key drug targets. The comparative modelling tool MODELLER was employed to integrate structural data from NMR spectroscopy, small-angle X-ray scattering and native mass spectrometry to generate a structural model of the receptor. The structural model was generated by assembling all of the individual domains of the structure as overlapping segments (Bugge et al., 2016).

Chapter 3

Modelling protein-protein interactions

3.1 Introduction

Protein-protein interactions (PPIs) are groups of two or more protein monomers that interact. These can be physical contacts or functional associations between proteins. This research has focused on PPIs which involve physical contacts between two or more proteins as a result of biochemical events. PPIs can be classified into homo-oligomeric complexes, if the PPI takes place between identical protein chains, or hetero-oligomeric complexes for PPIs that take place among non-identical protein chains with up to 70% of known complexes in the PDB being the former (Levy and Teichmann, 2013).

PPIs can be also be separated into permanent PPIs and transient PPIs. Permanent PPIs are usually associated with interactions between subunits of multi-subunit protein complexes. The protein monomers of these complexes are usually unstable on their own. Classical examples of stable PPIs include haemoglobin, core RNA polymerase and ATP synthase. By contrast, transient interactions, as the name implies, are temporary interactions between protein pairs that may be strong or weak. The binding of a transcription factor to the promoter region of a gene, signal transduction, binding between hormone receptors, antigen-antibody interactions and inhibition of proteases are examples of transient PPIs.

It has been speculated that there may only be 10,000 distinct protein interfaces (Aloy and Russell, 2004), and that the number currently characterised may be close to being complete (Gao and Skolnick, 2010), so that most of the known protein interfaces are probably reusable for modelling interactions in different contexts. Yet, some studies indicate that this is not entirely true: an analysis done using the Conserved Domain Database demonstrated that only 30% of protein domain families have a structural PPI representative suggesting that there may be many more interfaces to characterise (Goncearenco et al., 2014).

Protein interfaces usually consist of a solvent-excluded protein core region surrounded by a rim region that is partly buried. The protein core contributes to around 75% of the buried surface area, whereas the rim region contributes to the other 25%. The protein core and rim regions have different amino acid composition and conservation (Janin et al., 2008).

The characteristics of protein interfaces have been extensively studied using different datasets of protein complexes (Ansari and Helms, 2005; Bahadur et al., 2004; Chakrabarti and Janin, 2002; Janin et al., 2008; Jones and Thornton, 1996; Nooren and Thornton, 2003; Ofran and Rost, 2003). Transient protein interfaces tend to have a smaller interface region ($<$1500Å^2) than the permanent PPIs (with interface regions ranging from 1,500Å^2 to 10,000Å^2). Most protein interfaces are considerably flat, with planarity of about 2Å. Planarity were estimated by calculating the RMSD of all interface atoms from the least squares plane fitted through all interface atoms (Jones and Thornton, 1996). Permanent protein complexes have been shown to have more planar protein interfaces which are more closely packed and with less inter-subunit hydrogen bonds compared to non-obligate complexes (Jones and Thornton, 1996). Overall, protein interfaces are also more hydrophobic than the protein exterior (surface) but less than the inner region of the protein. Transient PPIs tend to be more polar, less hydrophobic than permanent PPIs (Ansari and Helms, 2005; Ofran and Rost, 2003).

Proteins interact through their interfaces and it is possible to define the interface regions (residues) from the 3D structure of the protein complexes. Protein interfaces can also be identified by the calculation of buried solvent accessible surface area (buried SASA) on complex formation. Interacting residues are defined as residues for which the total or side chain buried solvent accessible surface area is larger than 0Å^2-1Å^2 (Jones and Thornton, 1995). We can calculate the area by the following equation:

$$Buried\ SASA = sASA(protein\ 1) + sASA(protein\ 2) - sASA(protein\ complex) \tag{3.1}$$

In addition, various other definitions have been used to define interacting residues and are summarized in Table 3.1.

Table 3.1: Different definitions used to define interacting residues in the interfaces of protein complexes

Definitions	References
The distance between any of the atoms should be less than 5.0Å	Tsai et al. (1996)
The distance between any of their atoms, one from each, should be less than the sum of their corresponding Van der Waals radii plus 0.5Å	Tsai et al. (1996)
The distance between two sulfur atoms of a pair of cysteines, should be less than 2.56Å (two times the covalent radius of sulfur plus 0.5Å) (disulphide bonds)	Mosca et al. (2012)
The distance between all N-O and O-N atom pairs, should be less than 3.5Å (hydrogen bonds)	Mosca et al. (2012)
The distance between all N-O and O-N atom pairs, should be less than 5.5Å (salt bridges)	Mosca et al. (2012)

X-ray crystallography (X-ray), nuclear magnetic resonance spectroscopy (NMR) and cryo electron microscopy are the most commonly used experimental techniques used to determine the 3D structures of protein complexes. The two main protein complex 3D structure repositories are the PDB and PISA (Krissinel and Henrick, 2007). PDB contains all the author determined protein complexes, solved experimentally. By contrast, PISA provides predicted protein complex structures. Using graph theory, PISA enumerates all the protein complexes (biological assemblies) that can be potentially formed in a given biological unit. PISA also performs analysis of the chemical stability of each complex. Chemically stable complexes are identified as those with a positive free energy of dissociation (determined by the calculation of the free energy of binding and the entropy change term) (Krissinel and Henrick, 2007).

Databases such as 3D-complex (Levy et al., 2006) and DOCKGROUND (Douguet et al., 2006) were developed to reduce the structural redundancy of protein complexes found in the PDB. These databases serve as an important tool for biologists to understand the structure, dynamics and assembly of protein complexes that are found in nature and also for structural modelling of protein complexes (Ahnert et al., 2015; Marsh and Teichmann, 2015).

3D-complex uses graph theory to cluster complexes based on structural/sequence similarity and patterns of contacts between chains. All the chains are characterised using SCOP domain architectures (Murzin et al., 1995). In addition, manual inspection of complexes is used to detect erroneous PDB complexes. Examples of errors in complexes include: 1) protein complexes that fail to have symmetry 2) protein complexes that were assigned to an incorrect symmetry type 3) protein complexes that have wrongly assigned chain numbers 4) identical protein complexes that are assigned to a different quaternary structure by PDB and PQS (predecessor of PISA).

By contrast, the DOCKGROUND resource does not use graph theory to classify the protein complexes. Protein complexes are separated into pairwise complexes. Pairwise complexes are defined as pairs of interacting chains that belong to the same PDB entry. The SCOP domain definition is also used to characterise the chains in the pairwise complex. In addition to that, all the protein chains are annotated with AEROSPACI information that takes into account properties such as the protein structure's resolution, R-factor and stereochemical properties (Chandonia et al., 2004).

The DOCKGROUND resource also removes illegitimate complexes such as interwoven chains, tangled chains and chains with interacting unfolded termini parts. Interwoven chains generally constitute a single polymer which has been broken into 2 sections because of missing residues in the X-ray structure. A tangled chain is a free unfolded segment with more than 6 residues which interact exclusively with the neighbouring chain. Similar to tangled chains, chains with interacting unfolded termini have an unfolded termini segment which interacts with the neighbour chain. After removing illegitimate complexes, pairwise complexes are then clustered using structural (based on the SCOP's classification scheme) or sequence similarity. In summary, DOCK-

GROUND is a database of pairwise complexes. Separating complexes in this way is valuable because multimeric complexes are composed of many pairwise complexes.

ProtinDB (Jordan et al., 2011) is a comprehensive web server that allows the user to construct a dataset of protein-protein interface residues from the PDB. Users can select the different cut-offs used to characterise the interface residues (distance between C-α atoms, distance between atoms based on Van der Waals radius) and surface residues (relative accessible surface area threshold value). In addition, users have the option to extract the information from the biological unit file or asymmetric unit files from the PDB.

Protein chains are generally composed of smaller structural units called protein domains. These are compact, local, and semi-independent units of protein structure, which are capable of performing their own functions. Besides PPI databases (which are typically established as chain-chain interaction databases), there are databases that focus on domain-domain interactions (DDIs). DDI databases such as 3did (Mosca et al., 2013b) and iPfam (Finn et al., 2014) involve the projection of Pfam domains (Punta et al., 2012) onto 3D structures obtained from the PDB.

There are also PPI databases that include predicted PPI information. PrePPI database developed by the Honig Lab is a database that contains both predicted and experimentally determined PPIs. PrePPI predicts over 2 million PPIs including 31,402 PPIs for yeast and 317,813 PPIs for human (Zhang et al., 2013). NCBI-IBIS is another comprehensive predicted and experimental PPIs database. NCBI-IBIS contains not only PPIs, but also Protein-DNA, Protein-RNA and Peptide-Ion interactions (Shoemaker et al., 2012).

Interactome3D (I3D) is another major resource providing 3D annotated protein networks. I3D is based on an automatic pipeline that first collects experimentally determined PPIs from publicly available databases and uses this data to build protein networks. Subsequently, protein structure information from the Protein Data Bank (PDB) (Berman et al., 2000) is added onto the networks. If the pipeline fails to find a structure for a particular PPI, I3D tries to model the protein structure using homology models or domain-domain information (Mosca et al., 2012). Currently, I3D annotates about 14% of the 81,000 known human PPIs with protein structures.

Table 3.2 and Table 3.3 summarise some of the current PPI/DDI databases with structural information.

Table 3.2: PPI databases that contain structural information

Name	Source of data	Number of Entries	Strength	Last Update	References
3did	DDIs - Pfam	385,177 structures for 10,593 unique DDIs	Identifies and groups different binding modes by clustering similar interfaces into "interaction topologies". Includes a pipeline for the discovery and annotation of novel domain-motif interactions.	Jan 2017	**Patrick Aloy** Institute for Research in Biomedicine, Spain (Mosca et al., 2013a)
iPfam	DDIs - Pfam	278,678 structures, corresponding to 9,516 unique DDIs	Provides the details of domain interactions at residue and atomic resolution Presents visualizations for analyzing the interactions from both sequence and structure perspectives.	November 2014	**Robert D. Finn** The Wellcome Trust Genome Campus (Finn et al., 2014)
KBDOCK	DDIs - Pfam	239,494 DDIs	Provides structure-based homology docking templates Uses a spatial clustering algorithm to define domain family binding sites	June 2013	**Dave Ritchie** INRIA, LORIA, Campus Scientifique (Ghoorah et al., 2014)
GWIDD	PPIs - BIND, DIP	120,818 PPIs 10,924 experimental structures 14,635 predicted structures	Predicts model structures using homology docking	August 2009	**Ilya Vakser** University of Kansas (Kundrotas et al., 2010)

Table 3.3: PPI databases that contain structural information

Name	Source of data	Number of Entries	Strength	Last Update	References
NCBI IBIS	GenBank, PDB/MMDB, and CDD databases	Number of domains/chains with experimental PPIs: 220,808 Number of domains/chains with predicted PPIs: 277,256	Reports protein-protein, protein-small molecule, protein nucleic acids and protein-ion interactions observed in experimentally determined structures. Predicts PPIs and binding sites by inspecting the protein complexes formed by close homologs of query	January 2017	**Thomas Madej/ Anna Panchenko** NCBI (Shoemaker et al., 2012)
PrePPI	PPIs - MIPS, DIP, IntAct, MINT, HPRD and BioGRID	Experimental: 199,863 PPIs Predicted: ∼2 million PPIs	Predicts PPIs using an in-house method that combines structural and non-structural information (e.g. coexpression, functional similarity and evolutionary similarity) (Zhang et al., 2012) Provides crude structural interaction models for lots of interactions Assigns probability for each interaction using a Bayesian framework	January 2012	**Barry Honig** Univerisity of Columbia (Zhang et al., 2013)
Interactome3D	PPIs - Intact, MINT, DIP, MPIDB, MatrixDb, InnateDb, BioGRID, BIND and HPRD	Structural details for over 25,000 PPIs for 16 model organisms	3D annotated protein network database	January 2017	**Patrick Aloy** IRB, Spain (Mosca et al., 2012)

3.1.1 Structural modelling of protein complexes

Despite the significant effort incurred by the structure genomics initiatives to characterise protein complexes (Nair et al., 2009; Schwede, 2013; Terwilliger, 2011), the number of complexes solved is still relatively low (Mosca et al., 2012; Stein et al., 2011). Therefore the development of new/better protein complex structural modelling tools is crucial to get a better picture of protein complex space.

There are two different computational approaches to predict the 3D structure of bound protein complexes : (1) Template-free or docking and (2) Template-based methods.

3.1.1.1 Template-free modelling/docking

Protein structures of monomeric proteins in the complex, that have been solved experimentally or predicted, serve as the starting point for protein docking. Protein docking involves the assembly of two monomers in different binding orientations. Searching through 6-dimensional space (which covers both translational and rotational spaces) is usually computational expensive. During rigid-body docking, the internal geometry of the monomer structure such as bond angles, bond lengths and torsion angles are maintained. By contrast, flexible docking allows changes in internal geometry. Since there are millions of docking conformations available to consider, protein docking is usually very computationally expensive.

After thousands of docking poses have been sampled, the process of scoring and ranking the poses usually follows. Many scoring functions have been developed, but those used in current state-of-the-art methods are usually based on statistical potentials, molecular mechanics, binding affinity ($\Delta\Delta$G), or prediction of hotspot residues (Moal et al., 2013). (See Figure 3.1)

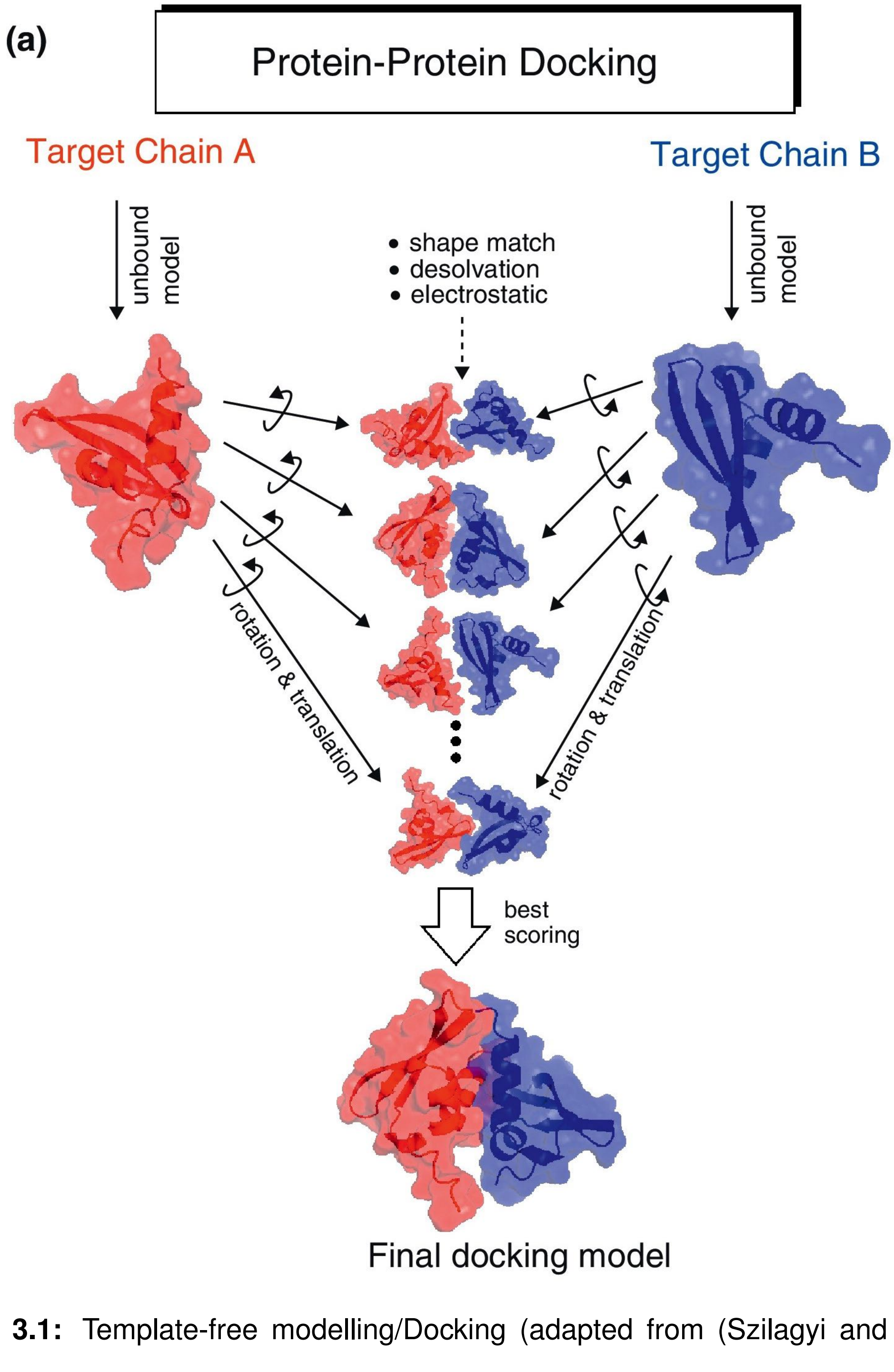


Figure 3.1: Template-free modelling/Docking (adapted from (Szilagyi and Zhang, 2014))

The current state-of-the-art template-free modelling methods were reviewed recently (Rodrigues and Bonvin, 2014). HADDOCK (Dominguez et al., 2003) and ClusPro (Comeau et al., 2004) were the two best performing groups in recent CAPRI protein complex prediction competitions (Lensink et al., 2016, 2017; Lensink and Wodak, 2013). HADDOCK uses water, buried surface area, interaction restraint energy and desolvation to sample and score the docking poses. Similarly, ClusPro uses electrostatics, interaction restraint energy and desolvation free energy for filtering models. (See Table 3.4, CAPRI will be explained in Section 3.1.2.1)

Table 3.4: Prediction performance of web servers in recent CAPRI rounds. For the second,third and fourth column, the first numbers are the number of predicted targets, the numbers inside the brackets are the number models that are of good quality. (Adapted from (Lensink et al., 2016, 2017; Lensink and Wodak, 2013)).

Server Group	Round 20-27	Round 30	Round 28-35
ClusPro (Comeau et al., 2004)	10 (6)	25 (16)	12 (5)
HADDOCK (Dominguez et al., 2003)	10 (4)	25 (16)	12 (2)
SWARMDOCK (Torchala et al., 2013)	4 (4)	25 (11)	12 (2)
GRAMM-X (Tovchigrechko and Vakser, 2006)	6 (0)	22 (6)	12 (1)
LZERD (Esquivel-Rodríguez et al., 2012)	2 (1)	25 (3)	12 (3)

A major weakness of protein docking is the significant loss in accuracy when trying to build protein complexes that involve a major conformation change (Moreira et al., 2010). New template-based approaches aim to address this problem.

3.1.1.2 Template based modelling

In contrast with template-free approaches, template based methods involve the construction of protein complexes by copying the structural architecture of related template complexes. Both sequence and structural similarity have been used to identify suitable templates.

Sequence based approach Sequence-based approaches are based on the assumption that similar sequences tend to share similar structures and functions and therefore similar binding orientations. If two query proteins are homologous to two other proteins that are known to form a complex, we can then use the known proteins as structural templates to model the query complexes. In 2003, Aloy et al. conducted a study to understand the relationship between sequence and interaction divergence. They demonstrated that pairs of proteins with sequence identity $>$40% tend to interact similarly. Inspired by this work, the Aloy group developed a structural modelling protocol and resource named Interactome3D (Mosca et al., 2012). Structural templates are identified by matching the query sequences of protein monomers with the sequences of protein complexes whose structures have been deposited in the PDB. Aloy et al. then modelled the protein complexes using the comparative modelling software MODELLER (Sali and Blundell, 1993).

To determine the accuracy of modelled PPIs, Aloy et al. performed a benchmark to model interactions for which experimental structures are available. The quality of the models was assessed using interface RMSD and the fraction of native residue-residue contacts (these criteria will be described in Section 3.1.2.1). The top-ranked models produced gave medium- to high- quality models for 52% of human interactions (obtained from experimentally determined PPI resources such as IntAct, MINT and HPRD), for which templates could be identified (Mosca et al., 2012).

Structure based approach (3D-template-based docking) Sometimes, proteins that are quite different in their sequences can share still high similarity in structure. Therefore structure based methods have also been developed. 3D-template-based docking is based solely on structural similarity (Kawabata, 2016; Kundrotas et al., 2012; Mukherjee and Zhang, 2011; Tyagi et al., 2012; Zhang et al., 2012). Structural similarity in theory should provide a better coverage because there are cases of structural and functional relationships that sequence similarity alone fails to identify. After suitable docking templates have been identified, individual components are superposed onto the docking templates. The resulting superposition is the structure of the predicted protein complex (See Figure 3.2).

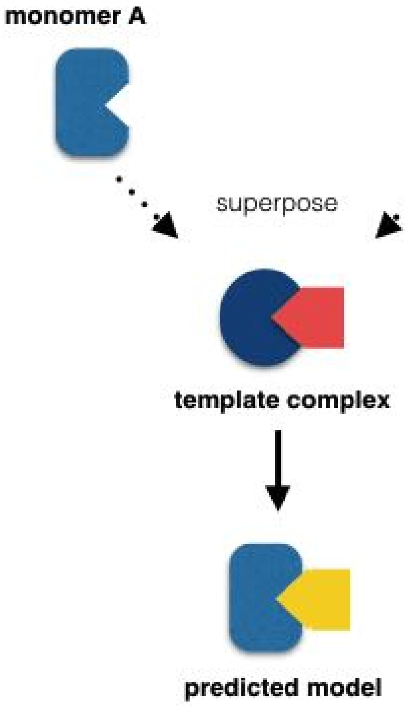


Figure 3.2: Superposing monomers onto the docking templates to produce a predicted protein complex.

PRISM (Kuzu et al., 2013; Tuncbag et al., 2012), VAKSER protocol (Kundrotas and Vakser, 2013) and Interactome3D (Negroni et al., 2014) are the current-state-of-the-art 3D-template-based docking pipelines that generate 3D models of complexes on a large scale. Table 3.5 compares the template pool, structural matching methods, filtering methods and ranking methods used by the three pipelines.

Table 3.5: Differences between PRISM, Interactome3D and Vakser pipelines

Methods	Template pool and matching methods	Filtering of docking poses	Ranking of docking poses
PRISM	7,922 non-redundant protein-protein interfaces from PDB/PISA Structural alignment tool: MultiProt (Shatsky et al., 2004)	40% of the template should structurally match the target surface (60% if the number of residues is less than 50) At least 15 interacting residues Maximal RMSD between matches is 2Å At least one hot spot should be matched if information is available	Global energy value (Calculated using Fiber-Dock (Mashiach et al., 2010))
Vakser method	11,932 non-redundant protein-protein interfaces Structural alignment tool: TM-align	Tm-score of at least one of the structural matches >0.4 At least 50% of the aligned residues should be on the surface At least 40% of the interface residues should be covered by the alignment	N/A Build all the models for a specific target that meet the selection criteria
Interactome3D	7,017 non-redundant protein-protein interfaces produced by DOCK-GROUND resource Structural alignment tool: TM-align	Less than 3 Cα-Cα clashes Mean accessible surface area by each chain lower than 250Å^2 Structural comparison of alignment: TM-score of 0.4 (Permissive) TM-score of 0.6 (Strict)	Value of the sum of the two TM-scores (one for each target monomer) obtained from the alignment between docking targets and docking templates

The application of 3D-template-based docking seems promising. The Vakser group suggest that in the near future is it is very likely that there will be docking templates available to model the human PPIs (i.e. human PPIs obtained from the DIP (Salwinski et al., 2004) and BIND (Alfarano et al., 2005) databases) as there are increasing numbers of structures in the PDB (Kundrotas et al., 2012). In order to evaluate the quality of the models built, Vakser et al. benchmarked their method using PDB protein complexes solved in 2009 to 2011, while using protein complexes solved before 2009 as docking templates. About one-third of all the models produced by them were of good quality (interface RMSD <5Å). (Interface RMSD will be explained in detail in the next section).

Recently, Negroni and co-workers undertook a similar analysis, but focused only on the applicability of template-based docking in the Twilight Zone (using templates with sequence identity less than 30%) (Negroni et al., 2014). The benchmark was done using the protein-protein docking benchmark version 4 dataset (PBD4) (Hwang et al., 2010). With a coverage of 92% and precision of 23% for the models built, the result was in agreement with the analysis conducted by the Vakser group. This suggests that the number of good quality models is still limited. Therefore, there is still a need to undertake further research into 3D-template-based docking.

3.1.2 Community-wide protein complex modelling assessment experiment

Inspired by CASP, the Critical Asssessment of PRedicted Interactions (CAPRI) community-wide experiment has been running since 2001 (Janin et al., 2003). CAPRI just held the 43th round of experiment with two to four rounds held every year. There are over 50 research groups worldwide that have participated in CAPRI. The main goal of CAPRI is to assess the quality of the predicted structures of protein complexes. A typical CAPRI round usually starts when the CAPRI organisers receive new protein complex structures from structural consortia. The amino-acid sequence of the target protein and some relevant information about the proteins is then provided to the

CAPRI participants. The participants are asked to model the 3D structure of the protein complexes. Any models submitted are assessed using the CAPRI assessment protocol (Janin et al., 2015; Méndez et al., 2003).

3.1.2.1 Assessing the performance of modelling methods

As with protein monomers, the best approach for assessing the performance of the methods is through structural comparison of the modelled complexes with the known complexes. The structural comparison scores that are usually considered include the fraction of native residue-residue contacts (Fnat), and the interface RMSD ($RMSD^{\text{interface}}$).

Fnat is defined as the number of native residue-residue contacts in the predicted complex divided by the number of residue contacts in the native complex. Residues are considered to be in contact if any of their atoms are within 5Å. Some researchers calculate the fraction of non-native contacts (Fnon-nat) as well. Fnon-nat is defined by as the number of non-native residue-residue contacts in the predicted complex divided by the number of residue contacts in the native complex. $RMSD^{\text{interface}}$ accounts only for the RMSD of interface residues in the predicted complex compared with the known complex. Protein interface residues are defined as those having at least one atom within 10Å of an atom on the other molecule in the target structures.

Table 3.6 explains the CAPRI criteria used for labelling the predicted protein complexes.

Table 3.6: CAPRI criteria of labelling protein complexes (Adapted from Méndez et al.)

Quality of predicted models	Fnat	$RMSD^{\text{interface}}$
High	$\geq$ 50 %	$\leq$1.0 Å
Medium	$\geq$ 30 %	1.0 Å $<$ x $\leq$ 2.0 Å
Acceptable	$\geq$ 10 %	2.0 Å $<$ x $\leq$ 4.0 Å
Incorrect	$<$ 10 %	

3.2 Objectives

In the previous chapter, we demonstrated the advantages of using FunFams in protein monomer modelling. In this chapter, we explore the usage of FunFams in protein complex modelling. We start by assessing the structural conservation of protein-protein interfaces within FunFams. Since none of the databases use CATH domain architecture primarily to classify protein complexes, we project the CATH domains onto 3D structures obtained from the PDB and produce a library of known DDIs annotated with FunFam information. After that, we assess the conservation of protein-protein interfaces in FunFams. Subsequently, we explore the usage of FunFams in protein complex modelling (sequence based template based modelling) with the help of MODELLER. We also explore the value of FunFams for producing a template library for 3D-template based docking and benchmark it against one of the state-of-the-art template libraries.

3.3 Materials and Methods

3.3.1 Investigating the conservation of protein-protein interfaces in CATH-Gene3D FunFams

3.3.1.1 Generating a library of CATH-Gene3D FunFam domain-domain interfaces

ProtinDB (Jordan et al., 2011) is a comprehensive repository that enables the user to extract protein-protein interface residues from protein complexes found in the PDB. We used this information to determine which residues were interacting in the PDB entry. The interface residues in our study were defined according to the following criteria: (1) The distance between any two atoms, one from each side, was below <5Å (2) The protein interface must be exposed to the surface, therefore, a relative accessible surface area of at least 5% was used (3) We extracted information only from the biological units.

We defined two CATH domains as interacting domains if there were more than 5 interacting residues (Finn et al., 2014; Mosca et al., 2013a; Winter et al., 2006). We considered two types of interactions (1) within the same chain (intra-chain) and (2) between different chains (inter-chain). Once all the DDIs were identified, we extracted the corresponding residues from the biological assembly files provided by the PDB. Then, we identified the FunFam in which the domain was classified.

3.3.1.2 Calculating the conservation of protein-protein interfaces within FunFams and between FunFams across a superfamily

We used the DDI library produced to explore the conservation of protein-protein interfaces within FunFams and between FunFams in the same superfamily. For within FunFam comparisons, we only compared pairs of DDIs that were involved in the same FunFam (i.e. involve interacting partners from the same FunFams). For between FunFam comparisons, we compared pairs of DDIs that involved the same superfamily (excluding those interactions that involved the same FunFam).

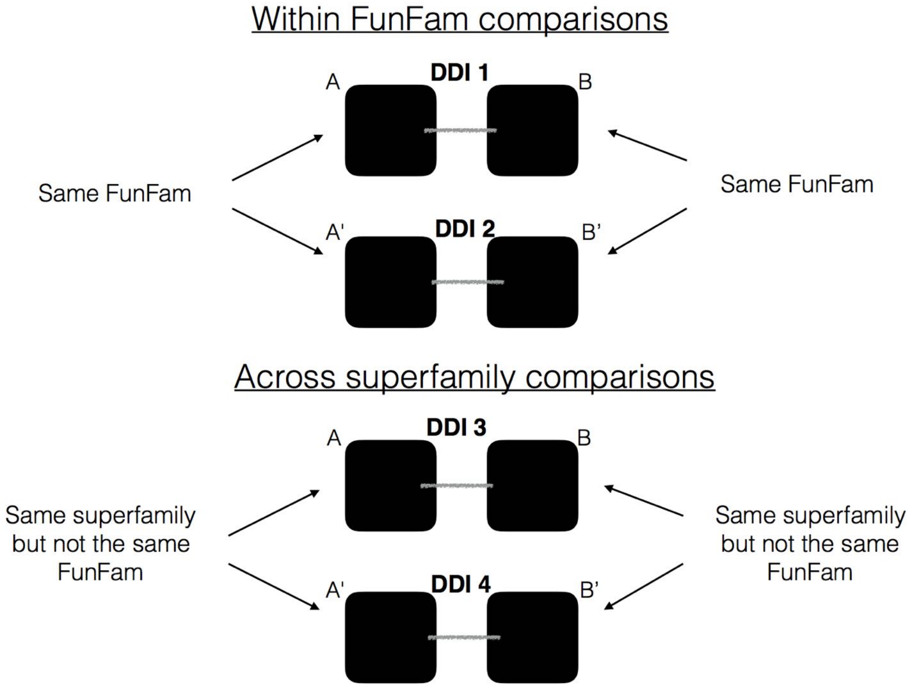

Within FunFam comparisons
A
DDI 1
B
Same FunFam
Same FunFam
A'
DDI 2
B'
Across superfamily comparisons
A
DDI 3
B
Same superfamily but not the same FunFam
Same superfamily but not the same FunFam
A'
DDI 4
B'

Figure 3.3: Within and between FunFam comparisons

Comparing all pairs of DDIs is very computationally expensive. Therefore, we clustered all the pairs of DDIs into 90% sequence identity clusters (S90 clusters). The structural representative for each S90 cluster was selected based on the X-ray resolution of the protein complex structure. Given two pairs of DDIs A-B and A'-B', where A-A' and B-B' are pairs of structures that belong to the same FunFams/Superfamily, interaction RMSD and interface RMSD were calculated. Interaction RMSD ($RMSD^{\text{global}}$) is a measure of the binding mode similarity, introduced by Aloy et al. (2003). It is the RMSD between a standard set of coordinates after two superpositions: one that optimally superimposes A' on A and one for B' on B. The standard 14 coordinates consist of the centres of mass for each domain plus six additional points defined by adding or subtracting 5Å to each of the x, y and z coordinates. A pair of interactions with a low $RMSD^{\text{global}}$ (e.g $<$10Å) are likely to have similar positioning of domains, but with a rotation of one domain relative to another. By contrast, Interface RMSD ($RMSD^{\text{interface}}$) measures the root-mean-square-displacement of the backbones of the interface residues. $RMSD^{\text{global}}$ is a global measures that accounts for both translational differences and domain rotations, while $RMSD^{\text{interface}}$ focus on the local changes of the protein interfaces.

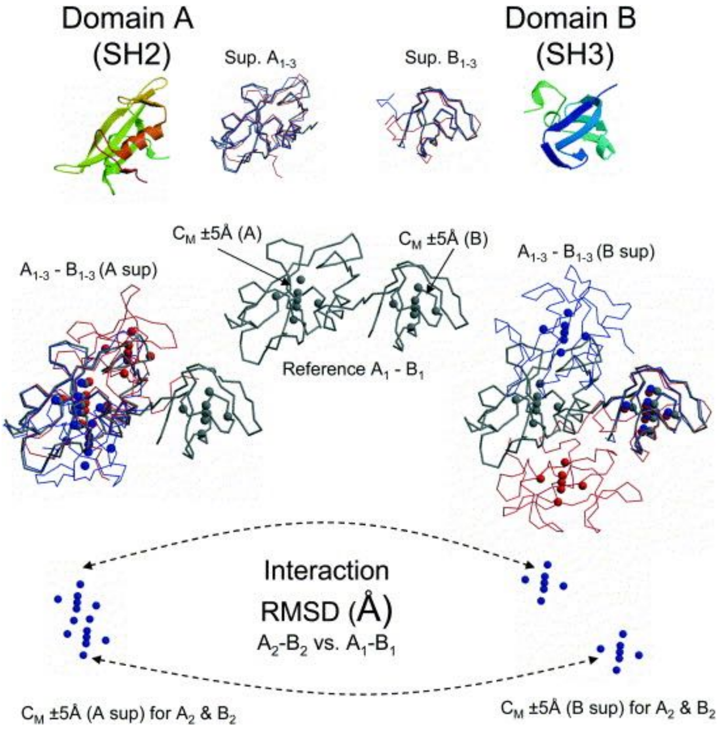


Figure 3.4: This figure illustrates how the $RMSD^{\text{global}}$ is calculated. Three complexes (1-3) are compared considering the first as a reference. Two superpositions were performed: one that optimally superimposes A' on A and one for B' on B. The spheres show the positions of the 14 atoms (i.e. centres of mass for each domain plus +/-5Å in each axis from the centres of mass to generate 6 further points.) following the superpositions. The RMSD of these positions were the $RMSD^{\text{global}}$. (Adapted from Aloy et al. 2003)

For each similarity (A and A' or B and B'), we also calculated the sequence identity and SSAP score (structural comparison obtained from the structural comparison tool SSAP(Taylor and Orengo, 1989)). The sequence identity is defined as the number of identical residues divided by the length of the smaller domain. For each pair of DDIs compared, there are thus two sequence and structural comparisons one for each monomer pair in the dimer. In subsequent plots, we used the minimum of the sequence identity and SSAP scores.

3.3.2 Modelling protein complexes - Template based modelling (MODELLER)

In this section, we explored the performance of FunFams in selecting templates for modelling binary PPIs using MODELLER. We performed two benchmarks to assess the accuracy of modelled PPIs built using FunFams to select templates, compared to (1) the template selection strategies used by Interactome3D (Mosca et al., 2012) (2) the template selection strategies using HHsearch (Söding, 2005) to identify templates. The protein complex modelling platform produced by MODELLER (Sali and Blundell, 1993) was used for both cases.

3.3.2.1 Complex modelling pipelines

FunFam pipeline Protein sequences of each query PPI were mapped into CATH-Gene3D domain sequences. All the query domain sequences were scanned for structural templates using FunFHMMer against the CATH FunFam HMM library. In order to build a model of the complex, template domain structures for a given query protein had to be from the same PDB entry. Therefore, selected FunFams are those that contain a domain from a PDB entry which contains all domains to be modelled and for which the E-value of the domain FunFam matches are $\leq$0.001. This threshold has been shown to give models for monomers.

After identifying the structural template to be used, we realigned the query sequence against the template's HHsearch HMM (The HMM is built by performing two iterations of HHsearch searches of the template against the UniProt20 database with an E-value threshold of 0.001) to obtain the query-template sequence alignment provided by HHsearch. (See Chapter 2 for more details).

We evaluated several methods for selecting templates. The assessment strategy is described below and the results are presented in Section 3.4.5.1. Subsequently, the comparative modelling software MODELLER was used to predict 10 models. Models with $<$5 residue-residue contacts were discarded. The top model was then selected based on the normalised DOPE score (Shen and Sali, 2006).

Interactome3D pipeline Since the Interactome3D pipeline is not publicly available, the results for Interactome3D were obtained from the Nature Methods' publication (Mosca et al., 2012). For this work, Intractome3D searched for structural templates using BLAST (Camacho et al., 2009) against the PDB BLAST library. The templates selected had a minimum sequence identity of 40% with the query targets. Structural templates with the highest sequence identity and coverage were selected. Sequence alignments between query and template were produced using MODELLER's built in alignment algorithm. Subsequently, MODELLER was employed to predict 5 models. Models with $<$5 residue-residue contacts were discarded. The top model was then selected based on the normalised DOPE score.

HHsearch pipeline All the query domain sequences were scanned against the CATH v4.0 HMM library for structural templates using HHsearch. As for the FunFam pipeline, template domain structures for the same query protein had to be from the same PDB entry. We selected the structural template with the highest probability to be a true positive provided by the HHsearch program. The structural templates must exceed the default probability score of 20 and have an E-value $\leq$0.001. We used the template-query sequence alignment provided by the HHsearch program.

We selected the best template based on the best method described in Section 3.4.4.1. Subsequently, the comparative modelling software MODELLER was used to predict 10 models. Models with $<$5 residue-residue contacts were discarded. The top model was then selected based on the normalised DOPE score.

3.3.2.2 Comparing different ways of selecting the best templates

We tested different combinations of scoring schemes (i.e. sequence identity, ZRANK global energy function score (Pierce and Weng, 2007), and the SOAP-PP protein interface statistical potential score (Dong et al., 2013)) to select the best templates for the modelled PPIs. ZRANK score is based on physics based potentials such as desolvation energy, Van der Waals, partials electrostatics (Pierce and Weng, 2007).

3.3.2.3 Benchmark dataset used in the comparison of the FunFam protocol with the Interactome3D protocol (Chain-chain interactions)

The quality of the Interactome3D models were obtained from the Nature Methods paper (Mosca et al., 2013a). Although, the benchmark target dataset is not publicly available, Aloy et al. reported that to benchmark Interactome3D they used all the experimentally solved PPIs from the November 2011 release of PDB. Subsequently, they selected those interactions for which a structural template could be found (with a minimum global sequence identity of 40%) as the targets.

We compiled two benchmark datasets in a similar manner to Interactome3D: (1) based on the November 2011 release of Interactome3D (2) based on the experimentally known complexes from the May 2015 release of Interactome3D. We selected PPIs for which domains in the query PPIs have been classified in CATH and structural templates could be found for both chains. Both the query datasets were chain-chain based and include both single-domain PPIs and multiple domain PPIs.

The size of benchmark target dataset built by the FunFam protocol and Interactome3D is different because the FunFam protocol covers only protein complexes which comprised domains that can be mapped onto CATH-Gene3D.

3.3.2.4 Benchmark dataset used in the HHsearch comparison (DDIs)

PBD5 (Vreven et al., 2015) is a non-redundant target protein complex dataset compiled by Zlab, which is the group that develops ZDOCK (Mintseris et al., 2007), one of the most robust template free docking methods in CAPRI. Based on the difference between the bound and the unbound form of the binding targets, all the docking targets were assigned an expected difficulty based on the $RMSD^{\text{interface}}$ and Fnat calculated after superposing the bound and the unbound forms of the known structure on top of each other. We classified the interactions into DDIs using the CATH domain definition v4.0. This gave 181 easy DDI targets and 103 medium/hard DDI targets.

Table 3.7: Criteria used to classify the difficulties of PBD targets

Difficulty	Criteria
Easy	Fnat $\geq$60% and $RMSD^{\text{interface}}$$\leq$1.5Å
Medium	[Fnat <60% and $RMSD^{\text{interface}}$$\leq$1.5Å] OR [1.5Å$<$$RMSD^{\text{interface}}$$\leq$2.2Å]
Hard	$RMSD^{\text{interface}}$ >2.2Å

3.3.3 Modelling protein complexes - 3D-Template-based docking

We developed a template based docking modelling pipeline adapted from the method developed by the Vakser group (Kundrotas and Vakser, 2013). Unlike the approaches described in Section 3.3.2, this approach uses structural comparison to find suitable templates. We performed two benchmarks to determine (1) how well can we recreate the Vakser pipeline (2) how well the template library based on CATH FunFams perform. We also explored methods to rank the docking poses. We used chain-chain PPIs in order to compare with Vakser. However, it is possible to adjust the pipeline to utilize domain-domain based PPI.

3.3.3.1 3D-Template based docking modelling pipeline

To model the structure of PPIs, the protein monomer structures for the PPI were obtained from the PDB. Subsequently, both monomers were structurally compared against an interface library of pairwise complexes described below. Structural alignment was done using TM-align (Zhang and Skolnick, 2005). TM-align was used because of its speed and the high accuracy of the alignment. Three filtering criteria were used to filter the docking templates: the TM structural similarity score of at least one of the structural matches (i.e. to one of the monomers in the complex) should be more than 0.4 (out of 1) to the template found in the library; the percentage of aligned residues located at the surface for both structures in the alignments should be $\geq$50%; At least 40% of the interacting residues should be covered by the alignment. Docking templates with a TM-score $>$0.9 were removed to prevent self-hits. Transformation matrices of the alignment were then used to produce the models. Figure 3.5

illustrates the pipeline.

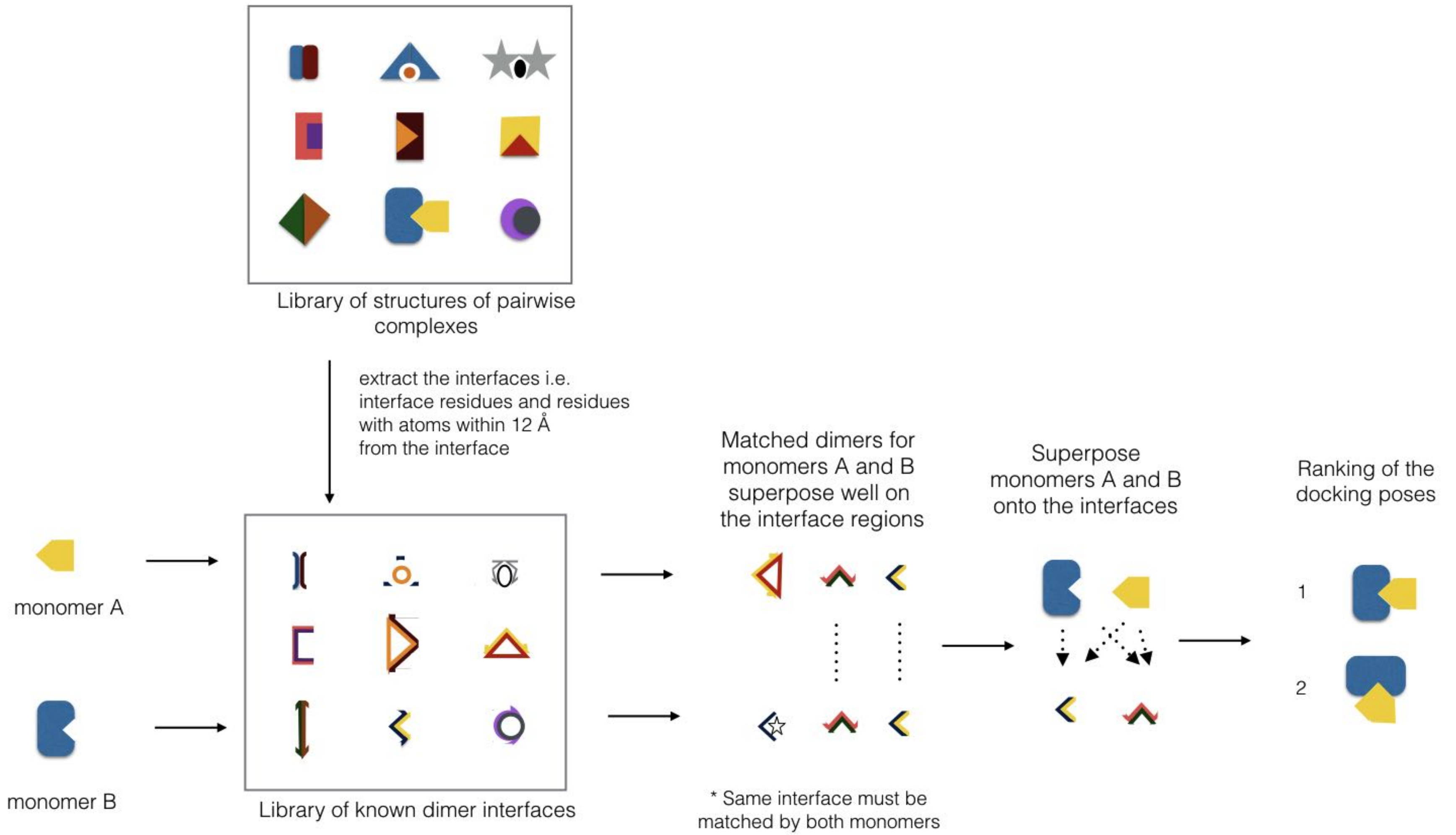

Library of structures of pairwise complexes
extract the interfaces i.e. interface residues and residues with atoms within 12 Å from the interface
monomer A
monomer B
Library of known dimer interfaces
Matched dimers for monomers A and B superpose well on the interface regions
* Same interface must be matched by both monomers
Superpose monomers A and B onto the interfaces
Ranking of the docking poses
1
2

Figure 3.5: 3D-Template based docking modelling pipeline

3.3.3.2 Template library of pairwise complexes

DOCKGROUND template library The only pre-built publicly available template library available was the DOCKGROUND template library (Douguet et al., 2006). This library was produced by the Vakser group and was built by starting with 12,134 protein complexes, solved with X-ray resolution of $\leq$3.5Å, mean accessible surface area buried by each chain of at least 250Å^2, and the number of interacting residues contributed from each chain of at least 10. After that, the protein-protein interfaces were extracted by selecting residues in the interface and within 12Å atomic distance from the interface (Sinha et al., 2012). The dataset was then clustered using a structural similarity TM-score cut-off of 0.9. This resulted in 7,107 non-redundant protein-protein interfaces of pairwise complexes. (See http://dockground.compbio.ku.edu for more details.)

The DOCKGROUND template library is based on the 2012 PDB release. We speculated using a newer PDB release would include additional structures who could improve the quality of the models built, as closer homologoues may be available as templates. We decided to build our own template library and use FunFams to cluster the templates.

Building the Lam library As with the DOCKGROUND template library, the Lam library was also composed of protein interfaces of pairwise complexes. It was adapted from the method developed by Douguet et al. (2006) for building a template library.

We started with 51,792 PDB entries by excluding all PDB entries, which were not protein entries, had X-ray resolution $\geq$3.5Å, and $<$30 residues in the protein chains. These PDB entries corresponded to 82,858 biological assembly files obtained from the PDB. The biological assembly file is the macromolecular assembly that has been shown to be the functional form of the molecule.

Biological assemblies were comprised of monomers, dimers, or multimers. Monomers were excluded from the benchmark. Dimers were composed of two interacting chains; no further processing was done. Multimers may be composed of a wide range of pairwise complexes depending on the nature of the particular proteins. For instance, there might be three types of pairwise complexes in a trimer (three chain multimer): A-B, A-C and B-C. The same theory applied to other multimers.

The next step was to make sure the pairwise complexes were interacting. Interacting pairwise complexes were defined as pairwise complexes with mean accessible surface area buried by each chain $\geq$250 and the number of interacting residues $\geq$10 (Douguet et al., 2006). Pairs of residues were considered as interacting if they are within 6 Å, to consider neighbouring residues, which constitute the scaffold of an interface (Kuzu et al., 2013). The accessible surface area was determined using NACCESS (Hubbard and Thornton, 1993). The calculation of mean accessible area buried by each chain was calculated using the following formula:

$$Mean\,buried\,SASA = \frac{sASA(protein\ \ 1) + sASA(protein\ \ 2) - sASA(protein\ \ complex)}{2} \tag{3.2}$$

After the filtering process, sequence and structural similarities are typically used to reduce the redundancy of the protein complexes (Douguet et al., 2006; Levy et al., 2006). Functional families have been shown to be functionally and structurally coherent (Das et al., 2015). Therefore, in constructing this dataset we explored the use of functional families to cluster the protein complexes. We grouped all the pairwise complexes based on the multiple domain architecture of FunFams (FF_MDA). FF_MDA used the functional family as the assignment unit and extends the assignment to both protein chains. Figure 3.6 illustrates the clustering process.

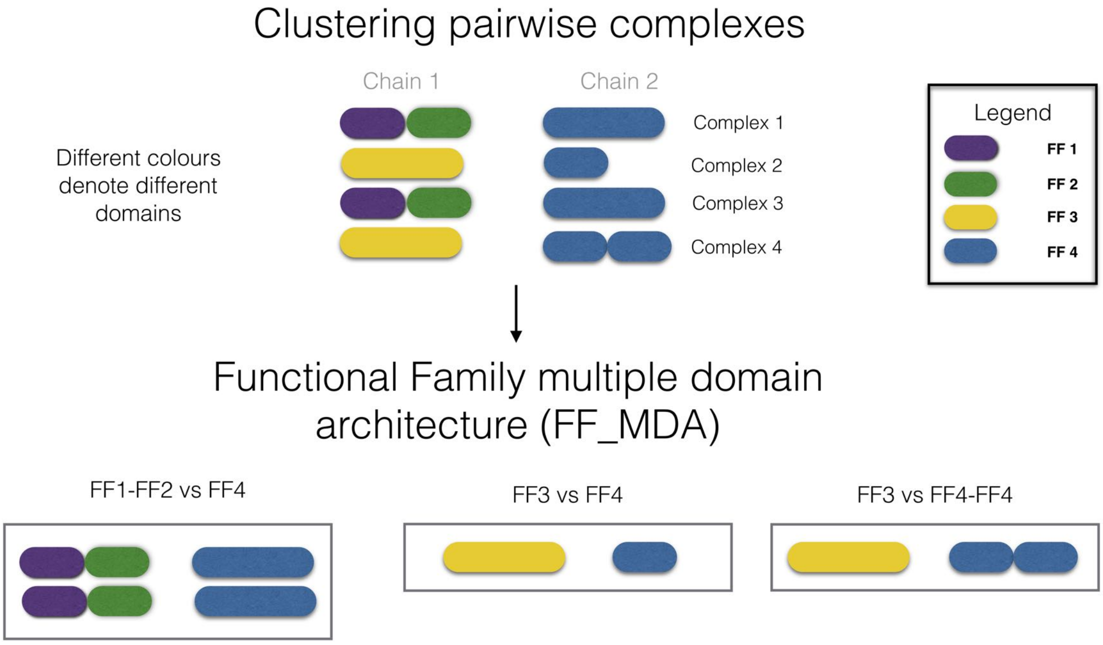

Clustering pairwise complexes
Chain 1
Chain 2
Different colours denote different domains
Complex 1
Complex 2
Complex 3
Complex 4
Legend
FF 1
FF 2
FF 3
FF 4
Functional Family multiple domain architecture (FF_MDA)
FF1-FF2 vs FF4
FF3 vs FF4
FF3 vs FF4-FF4

Figure 3.6: Chain-chain based protein complexes are separated into different subgroups based on the FunFam multiple domain architecture.

After the structural clusters were produced, the next step was to select the structural representative for each FF_MDA. The pairwise complexes with the lowest X-ray resolution were selected. If there was more than one pairwise complex for a particular PDB entry, the pairwise complex with the largest number of chains in the complex was selected. After that, the protein interfaces of the pairwise complexes were defined by selecting residues with atoms within 12Å from the interface (Sinha et al., 2012). Figure 3.7 summarises the template library building process.

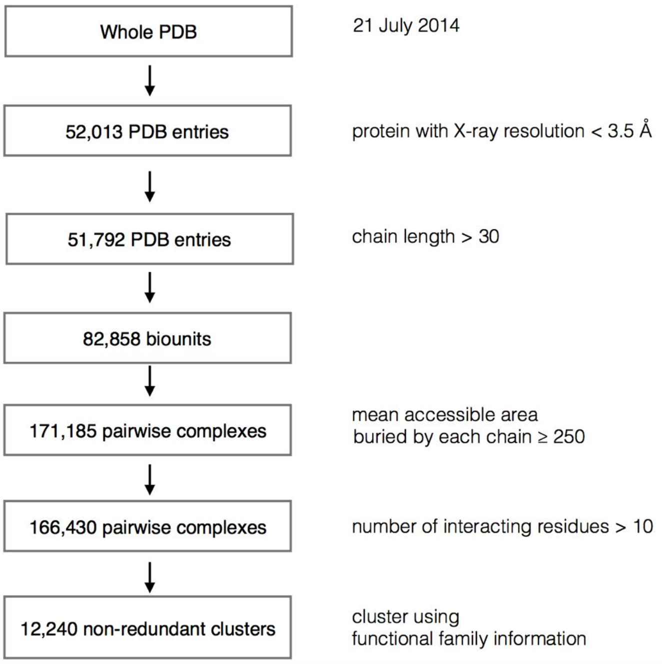

Whole PDB
21 July 2014
52,013 PDB entries
protein with X-ray resolution < 3.5 Å
51,792 PDB entries
chain length > 30
82,858 biounits
171,185 pairwise complexes
mean accessible area
buried by each chain ≥ 250
166,430 pairwise complexes
number of interacting residues > 10
12,240 non-redundant clusters
cluster using
functional family information

Figure 3.7: Summary of template library building process

3.3.3.3 Benchmarking the 3D-template based docking pipeline

We performed the first benchmark to determine how well we could recreate the Vakser pipeline using the publicly available DOCKGROUND template library. We used the same benchmark target dataset as the Vakser study (Kundrotas and Vakser, 2013). The benchmark dataset was composed of 372 non-redundant pairwise complexes clustered from the DOCKGROUND resource.

3.3.3.4 Benchmarking the 3D-template based docking pipeline using the Lam library

We performed another benchmark to assess the performance of the Lam library in template selection. We used the PBD4 (Hwang et al., 2010) non-redundant target protein complex dataset compiled by Zlab. This complex dataset comprises 80 easy cases, 22 medium cases and 17 difficult cases. This benchmark was included as it is now the most widely used benchmark for validating the predictions of protein complexes.

3.3.3.5 Ranking the docking poses

We tested different scoring schemes: (1) sum of the TM-scores (one for each target monomer) obtained from the alignment between docking targets and docking templates (Negroni et al., 2014), (2) MODELLER's normalised DOPE statistical potential score (Shen and Sali, 2006), (3) removal of steric clashes, and (4) MM-align (Mukherjee and Zhang, 2009) (Explained in the Introduction chapter) to rank the docking poses built.

3.3.4 Assessing the quality of models

To benchmark our MODELLER and 3D-template based docking pipelines, the models were assessed by doing a structural comparison with the known protein complexes. We used the assessment criteria adopted by CAPRI (Méndez et al., 2003). In short, a model was classified as good/bad quality according to the $RMSD^{\text{interface}}$ and the Fnat in the known protein complex that were also present in the model . See table below:

Table 3.8: Classifying the models into different categories

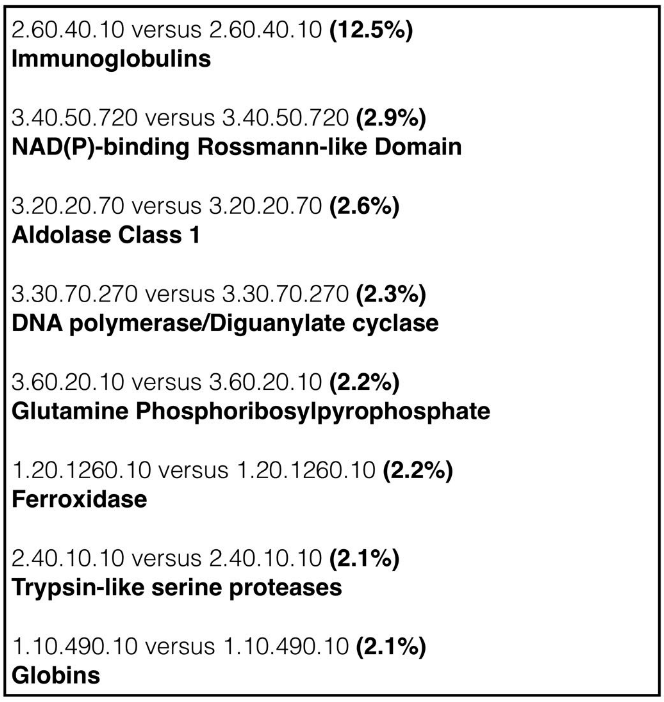
Quality of predicted models	Fnat	$RMSD^{\text{interface}}$
High	$\geq$50%	$\leq$1.0Å
Medium	$\geq$30%	1.0Å$<$x$\leq$2.0Å
Acceptable	$\geq$10%	2.0Å$<$x$\leq$4.0Å

3.4 Results and Discussion

3.4.1 Generating a library of CATH-Gene3D FunFam domain-domain interfaces

We obtained a total of 130,173 biological DDIs from the PDB using the protocol described in Section 2.3.2.1. 42% of the DDIs are intra-chain DDIs, and 58% of them are inter-chain DDIs. All the DDIs were grouped based on FunFam compositions. HomoDDIs are DDIs that involve the same pair of FunFam whereas heteroDDIs are

DDIs that involve different pairs of FunFams. 48% of the DDIs are heteroDDIs. Figure 3.8 and 3.9 summarise the most prevalent homoDDIs and heteroDDIs in our dataset. The most prevalent homoDDI is the immunoglobulin superfamily, followed by the classic Rossman fold superfamily, which contains domains that bind the enzyme cofactor NAD(P). The list also includes many enzyme superfamilies, such as Aldolase, Kinase, Ferroxidase.



2.60.40.10 versus 2.60.40.10 **(12.5%)**
Immunoglobulins

3.40.50.720 versus 3.40.50.720 **(2.9%)**
NAD(P)-binding Rossmann-like Domain

3.20.20.70 versus 3.20.20.70 **(2.6%)**
Aldolase Class 1

3.30.70.270 versus 3.30.70.270 **(2.3%)**
DNA polymerase/Diguanylate cyclase

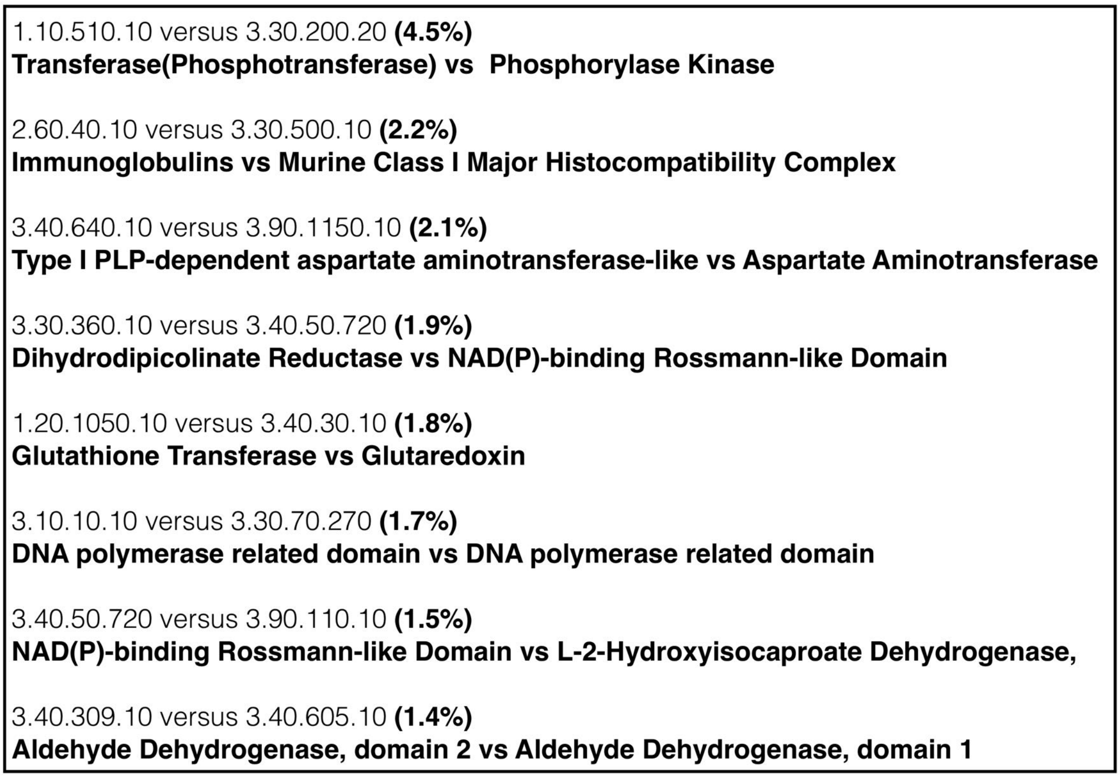
3.60.20.10 versus 3.60.20.10 **(2.2%)**
Glutamine Phosphoribosylpyrophosphate

1.20.1260.10 versus 1.20.1260.10 **(2.2%)**
Ferroxidase

2.40.10.10 versus 2.40.10.10 **(2.1%)**
Trypsin-like serine proteases

1.10.490.10 versus 1.10.490.10 **(2.1%)**
Globins

Figure 3.8: The most prevalent homoDDIs in the dataset. HomoDDIs are DDIs that involved the same CATH superfamily (e.g. superfamily 2.60.40.10 versus superfamily 2.60.40.10). We included the name of the name of CATH superfamily. We also included the proportions of domain that fall into these superfamilies.



1.10.510.10 versus 3.30.200.20 **(4.5%)**
Transferase(Phosphotransferase) vs Phosphorylase Kinase

2.60.40.10 versus 3.30.500.10 **(2.2%)**
Immunoglobulins vs Murine Class I Major Histocompatibility Complex

3.40.640.10 versus 3.90.1150.10 **(2.1%)**
Type I PLP-dependent aspartate aminotransferase-like vs Aspartate Aminotransferase

3.30.360.10 versus 3.40.50.720 **(1.9%)**
Dihydrodipicolinate Reductase vs NAD(P)-binding Rossmann-like Domain

1.20.1050.10 versus 3.40.30.10 **(1.8%)**
Glutathione Transferase vs Glutaredoxin

3.10.10.10 versus 3.30.70.270 **(1.7%)**
DNA polymerase related domain vs DNA polymerase related domain

3.40.50.720 versus 3.90.110.10 **(1.5%)**
NAD(P)-binding Rossmann-like Domain vs L-2-Hydroxyisocaproate Dehydrogenase,

3.40.309.10 versus 3.40.605.10 **(1.4%)**
Aldehyde Dehydrogenase, domain 2 vs Aldehyde Dehydrogenase, domain 1

Figure 3.9: The most prevalent heteroDDIs in the dataset. HeteroDDIs are DDIs that involved different CATH superfamilies (e.g. superfamily 1.10.510.10 versus superfamily 3.30.200.20). We also included the name of the name of CATH superfamily. We also included the proportions of domain that fall into these superfamilies.

3.4.2 Conservation of domain-domain interfaces within FunFams

We used the FunFam domain-domain interface library to explore the conservation of domain-domain interfaces within FunFams.

3.4.2.1 The relationship between the sequence similarity of FunFam relatives and conservation of the interfaces

Figure 3.10 shows the plot of conservation of domain-domain interfaces within FunFams, across a superfamily ($RMSD^{\text{interface}}$ versus sequence identity) with the 80th and 90th percentiles (i.e. 80% of 90% of the data are below the curve) marked.

Interactions are well conserved within FunFams (with $RMSD^{\text{interface}}$ $\leq$5Å) when the sequence identity is above 40%. Domain interfaces of intra-chain interactions are highly conserved with $RMSD^{\text{interface}}$ ≤2Å even for remote relatives (<30% sequence identity). The domain interfaces of inter-chain interactions are less conserved than the intra-chain interactions but they are still somewhat conserved (with $RMSD^{\text{interface}}$≤5Å) when the sequence identity is above 40%.

We also examined the conservation of interfaces across a superfamily (i.e. by comparing interfaces between representatives of different FunFams within the same superfamily). These are much less conserved than within FunFams. Again, as with comparisons within FunFams, the domain interfaces of intra-chain interactions are better conserved than inter-chain interactions.

Although inter-chain interfaces both within FunFams and across the superfamily tend to be conserved when the sequence identity is above 40%, there are significantly more inter-chain interfaces that are conserved (with $RMSD^{\text{interface}}$≤10Å), even at low sequence identity (i.e. <40% sequence identity) within FunFams than across the superfamily. (See Figure 3.10).

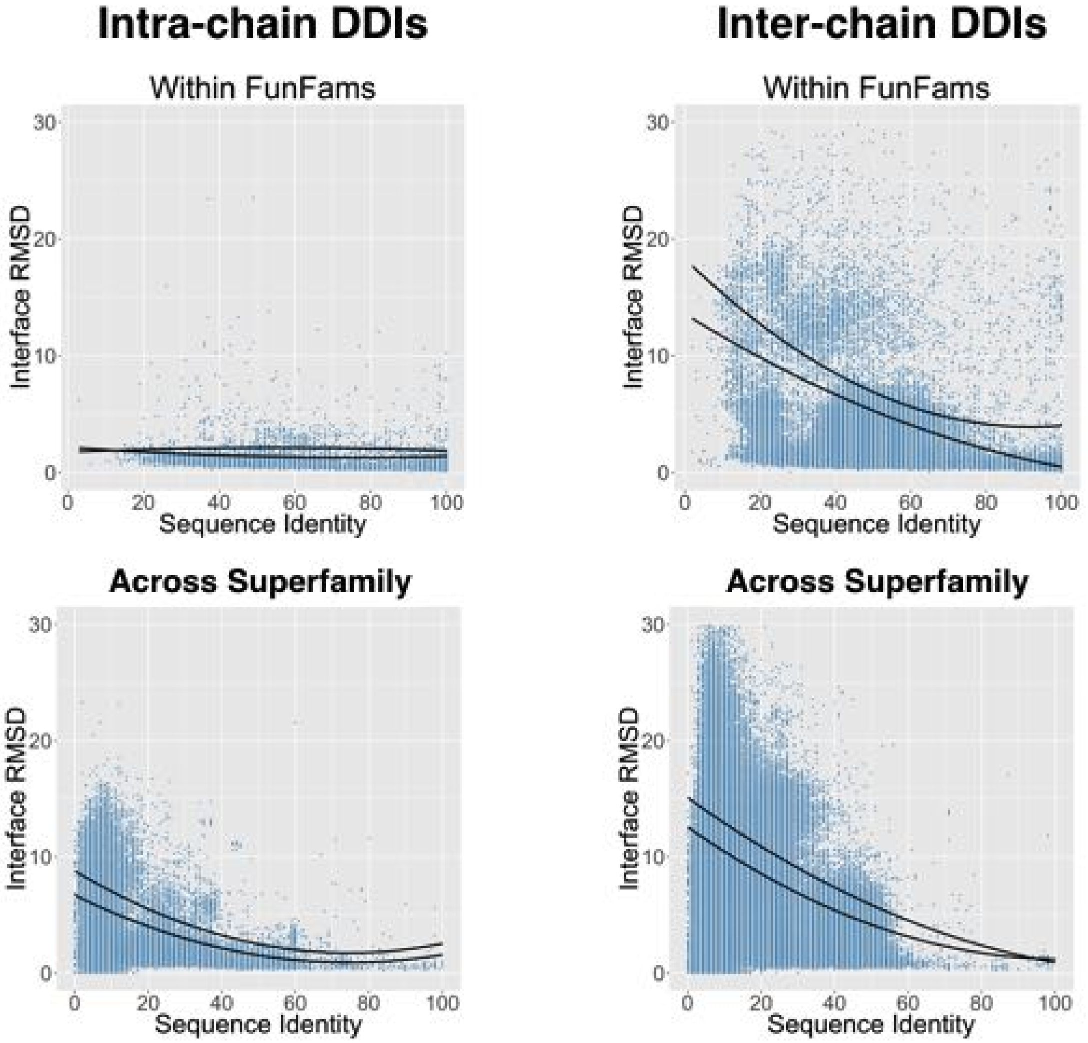


Figure 3.10: Relationship between sequence identity of domain relatives from FunFam/Superfamily and the conservation of their interface using $RMSD^{\text{interface}}$.The curves shows the 80th and 90th percentiles. Interactions were split into inter-chain and intra-chain.

Recent analysis in our group demonstrated that that some of the FunFams (4,100 out of the 101,000 FunFams) contain relatives that are annotated with diverse GO terms. These polluted FunFams are usually within the more functionally diverse superfamilies. We removed all the comparisons for these polluted FunFams as we hypothesized that functionally diverse relatives would be likely to differ in complexes and interfaces involved. For the non-polluted FunFams, we observed that 90% of the interfaces are well conserved ($RMSD^{\text{interface}}$ below 5Å) even for remote relatives (<30%

sequence identity) for both inter-chain and intra-chain interactions. We also used $RMSD^{\text{global}}$ to quantify the conservation. 80% of the interactions within non-polluted FunFams are well conserved (with $RMSD^{\text{global}}$ below 10Å). (See Figure 3.11 and 3.12)

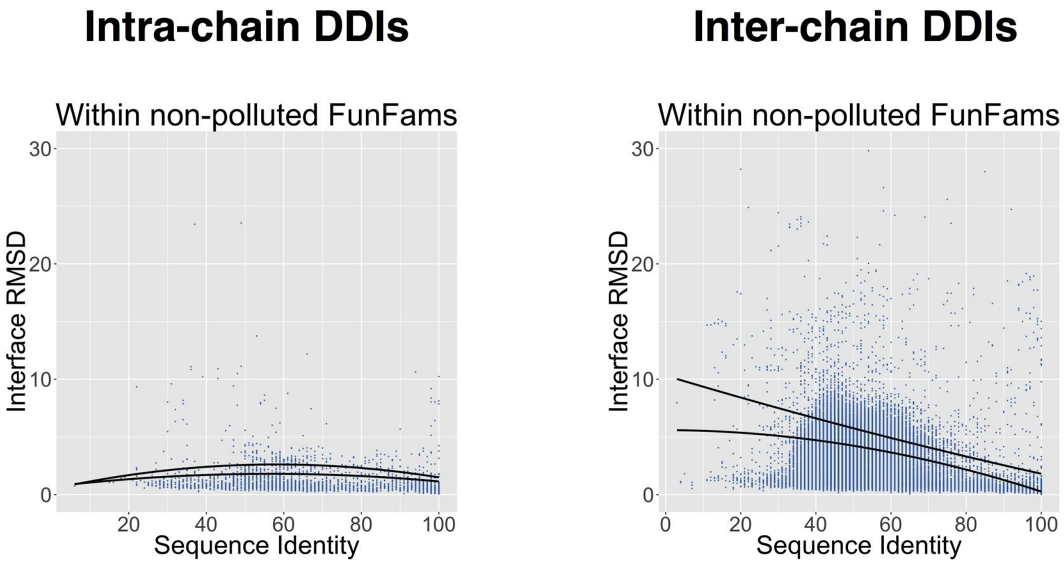


Figure 3.11: Relationship between sequence identity of domain relatives from non-polluted FunFams and the conservation of their interface using $RMSD^{\text{interface}}$.The curves shows the 80th and 90th percentiles. Interactions were split into inter-chain and intra-chain.

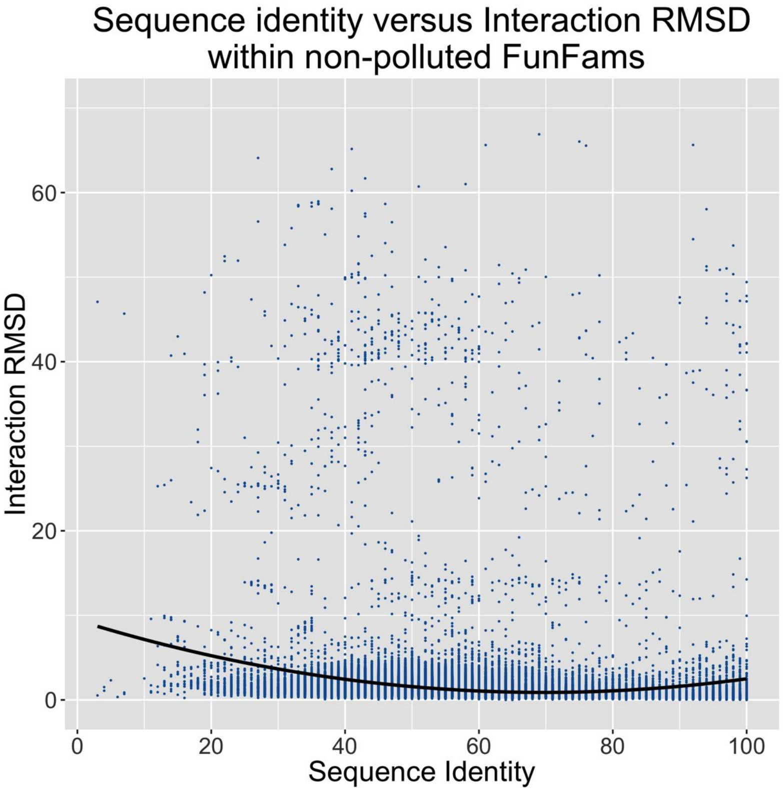


Figure 3.12: Relationship between sequence identity of domain relatives within non-polluted FunFams and conservation of their interface using $RMSD^{\text{global}}$.The curves shows the 80th percentile. This plot includes both inter-chain and intra-chain interfaces.

3.4.2.2 The relationship between the structural similarity of FunFam relatives and conservation of the interfaces

A similar trend was observed for the relationship between the structural similarity of the domain relatives in the FunFams and the conservation of the interface by $RMSD^{\text{interface}}$. We used SSAP to quantify the structural similarity. SSAP score is range from 0 to 100, with a score of 100 being very structurally similar. Overall, there's a higher conservation for intra-chain interactions than inter-chain interactions. Intra-chain domain interfaces are well conserved even for remote relatives within FunFams. The domain interfaces of inter-chain interactions are well conserved (with $RMSD^{\text{interface}}$ around 5Å) when the SSAP score is above 90 for both within FunFam and across superfamily comparisons. For lower values of SSAP, the inter-chain interfaces are not so well conserved. For non-polluted FunFam comparisons, 90% of the interfaces tend to be well conserved for both inter-chain and intra-chain interactions (See Figure 3.13). We also used $RMSD^{\text{global}}$ to quantify the conservation. 80% of the interactions within non-polluted FunFams are well conserved (with $RMSD^{\text{global}}$ below 10Å). (See Figure 2.14)

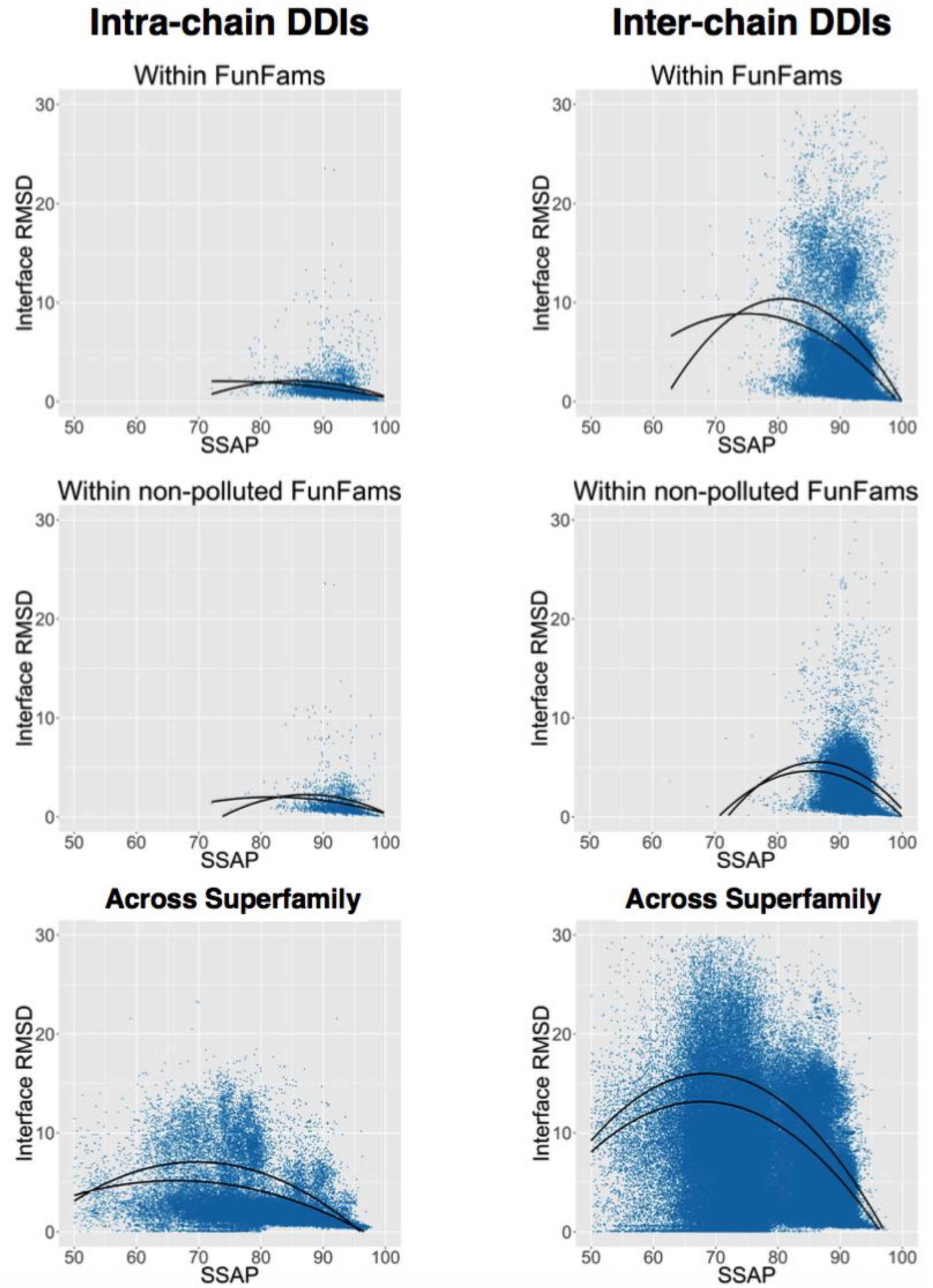


Figure 3.13: Relationship between structural similarity of domain relatives from FunFam/Superfamily and conservation of their interface using $RMSD^{\text{interface}}$. The curves shows the 80th and 90th percentiles. Interactions were split into inter-chain and intra-chain.

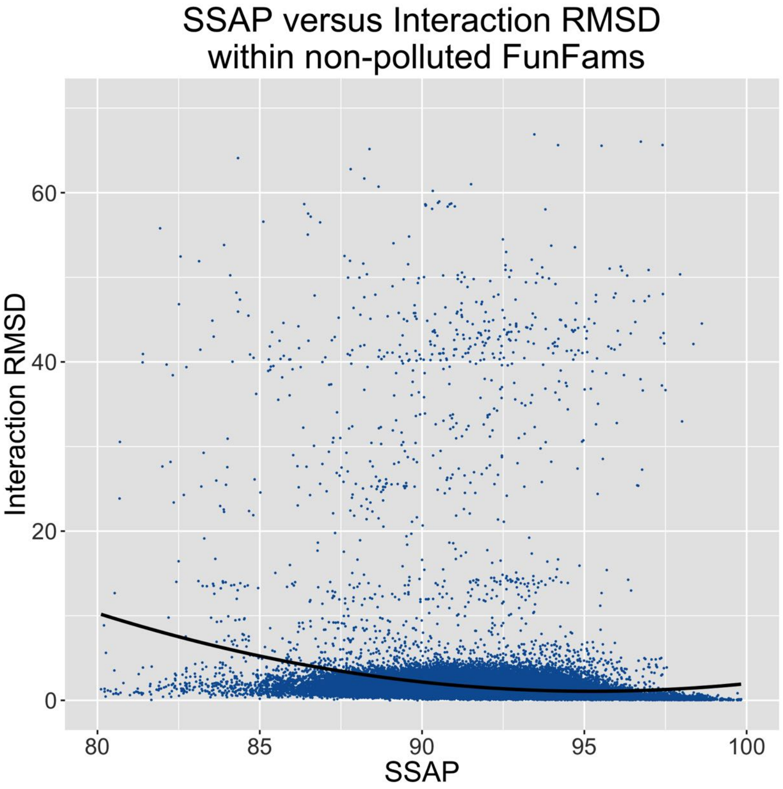


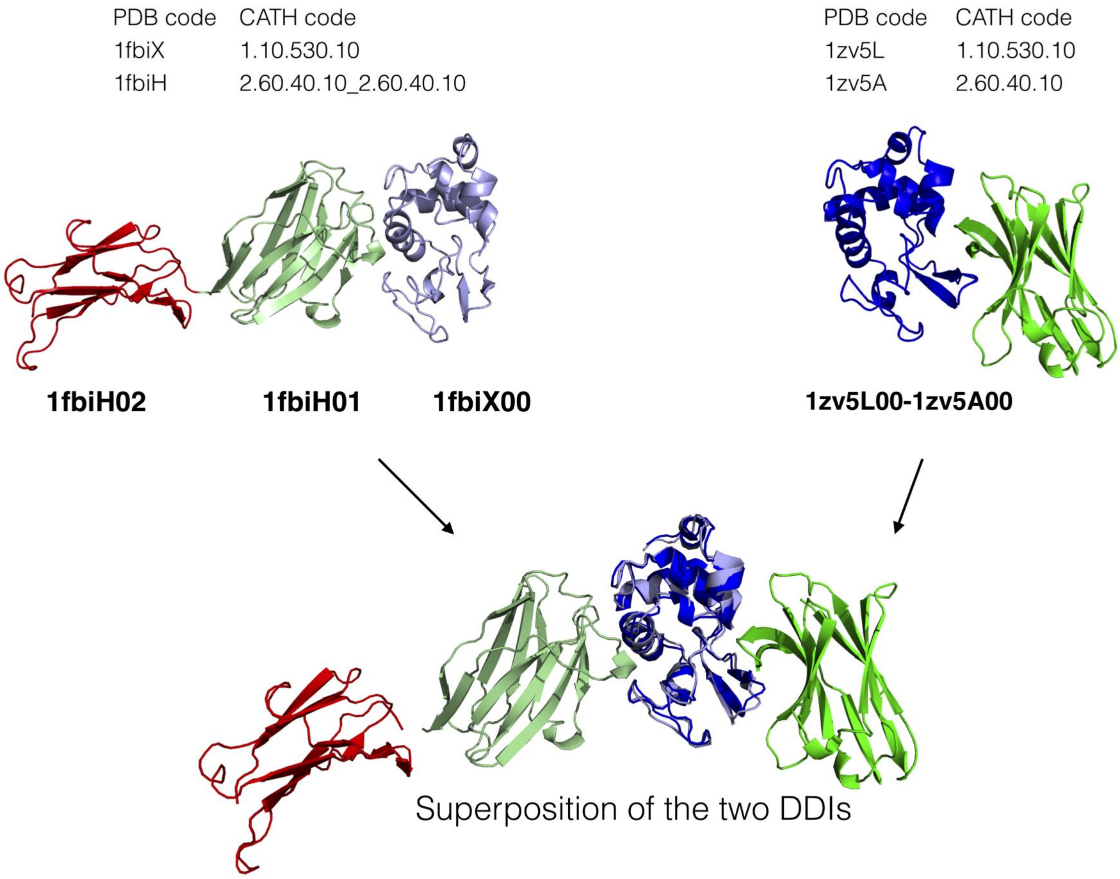
Figure 3.14: Relationship between structural similarity of domain relatives within non-polluted FunFams and conservation of their interface using $RMSD^{\text{global}}$.The curve shows the 80th percentile. This plot includes both inter-chain and intra-chain interfaces.

In summary, our analysis demonstrates that domain-domain interfaces tend to be well conserved within FunFams. However, some FunFams are polluted with relatives that have somewhat different functions and interfaces. In the following section, we investigate some examples of these polluted FunFams further.

3.4.3 Analysis of variation in complex structures in the polluted FunFams

3.4.3.1 Different multiple domain architectures (MDAs)

We analysed some extreme outliers from our plots (ie in figure 3.10 and figure 3.13) and found that there were differences in the MDA in these outlying pairs. See Figure 3.15 for an example where two pairs of DDIs are from chains with different MDA.



PDB code	CATH code
1fbiX	1.10.530.10
1fbiH	2.60.40.10_2.60.40.10

PDB code	CATH code
1zv5L	1.10.530.10
1zv5A	2.60.40.10

Figure 3.15: Both DDI 1fbiH01-1fbiX00 and DDI 1zv5L00-1zv5A00 were originated from chains having different MDAs. They involved the same FunFam pairs which involves the interaction between lysozyme (blue domain) and antibody (green domain). We speculate that the different in interface region is due to the fact that different antibody recognising the different part of antigen.

3.4.3.2 Different oligomeric state

DDI 3o7mA00-3o7mB00 and DDI 1tc1A00-1tc1B00 are both crystal structures of hypoxanthine phosphoribosyltransferase. DDI 3o7mA00-3o7mB00 is from *Bacillus anthracis*, while DDI 1tc1A00-1tc1B00 is found in *Trypanosoma cruzi*. Both of the DDIs are homodimers, so all the domains are in the same FunFam. They are highly similar in structure with a SSAP score (between the most dissimilar domains) of 92 and a sequence similarity (between the most dissimilar domains) of 40%.

The PDB entry of 3o7m is a tetramer and 1tc1 is a dimer. In the 3o7m tetramer, two of the domains (the blue(A”) and the violet (A”’)) have the same structure and interface as DDI 1tc1A00-1tc1B00. However the green (A') and the blue (A") domains have different interface. We speculate that the dimer 1tc1A00-1tc1B00 may belong to a larger interaction complex or the hypoxanthine phosphoribosyltransferase complex found in *Bacillus anthracis* has undergone duplication and formed a tetramer. The comparison of pairs for these DDIs in the FunFam involves comparing all pairs shown in 3.16 (e.g. A' A" versus A"A"'). However, as can be seen from the figure these will have different interfaces.

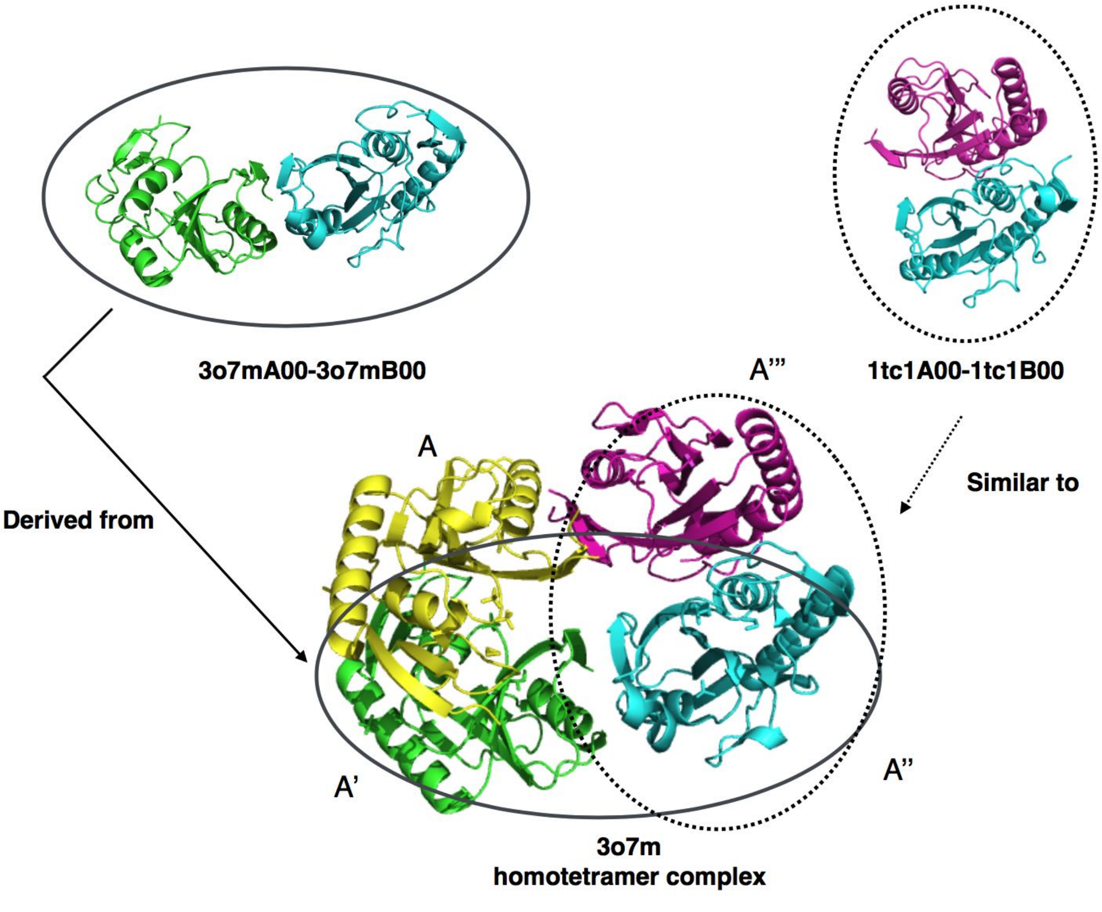


Figure 3.16: Two homo DDIs that are part of a larger interaction complex. Similar DDIs are found as as a dimer in 1tc1 and a tetramer in 3o7m

We colored the DDI comparisons within polluted FunFams according the similarity in the oligomeric state (See Figure 3.17). Some of the cases of DDI comparison with high $RMSD^{\text{interface}}$ did shares different oligomeric states.

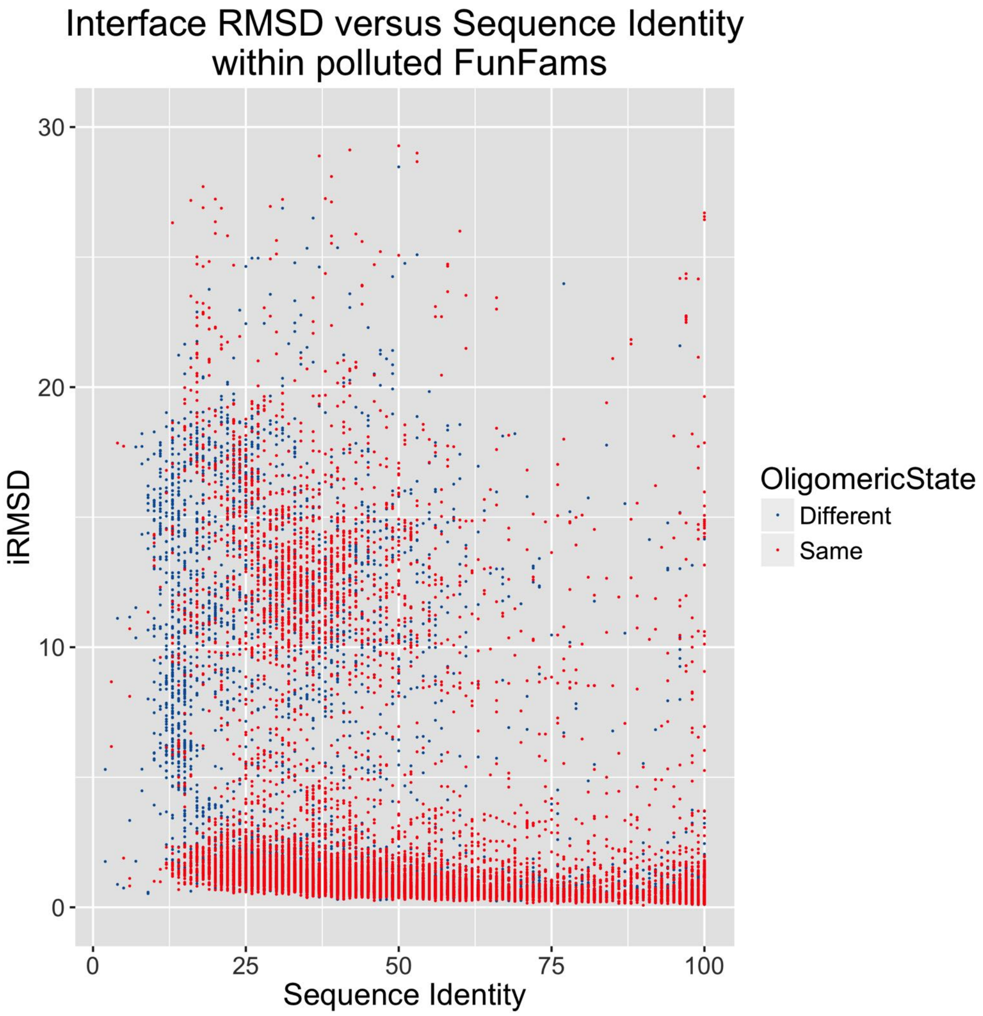


Figure 3.17: Relationship between sequence identity of domain relatives within polluted FunFams and conservaton of their interface using $RMSD^{\text{interface}}$. DDIs in the same (red) or different (blue) oligomeric states were colored accordingly.

3.4.3.3 Other possible reasons

Below are some other possible reasons explaining why interactions may be different within FunFams:

- One of the DDIs may have crystal packing/protein complex symmetry issues.
- One or more of the domains in the complex are in an active/inactive states.
- DDIs in a pair are involved in different cellular locations - where they are involved in different assemblies.

3.4.4 Assessing the performance of FunFams in selecting templates for modelling protein complexes using MODELLER

In this section, we report the performance of FunFams for selecting templates to model binary PPIs using MODELLER. We performed two benchmarks to assess the accuracy of modelled PPIs built using FunFams to select templates (1) comparing against the template selection strategies used by Interactome3D (Mosca et al., 2012) (2) comparing against the template selection strategies using HHsearch (Söding, 2005) to identify templates. The protein complex modelling platform MODELLER (Sali and Blundell, 1993) was used in both cases.

3.4.4.1 FunFam protocol for selecting the best template

Interactome3D identifies structural templates by BLAST and then selects the best templates based on the highest global sequence identity match. The FunFam protocol scans the query sequence against the FunFam library and then selects a relative with a known structure from the best matched FunFam.

We investigated various ways of choosing the best templates from the FunFams for the FunFam protocol. As well as using the global sequence identity, we also tested ZRANK global energy score and SOAP-PP protein interface statistical potential to select the best template. The best approach was determined by assessing the quality of the models by superposing them on the native structures. Figure 3.18 illustrates the quality of models selected using different combinations of scoring schemes. Only the top ranked models are shown in the plot. Sequence identity between query and template, ZRANK global energy score and SOAP-PP protein interface statistical potential were combined in different ways. The best method was the protocol that used only sequence identity. This protocol selected the highest number of good quality models.

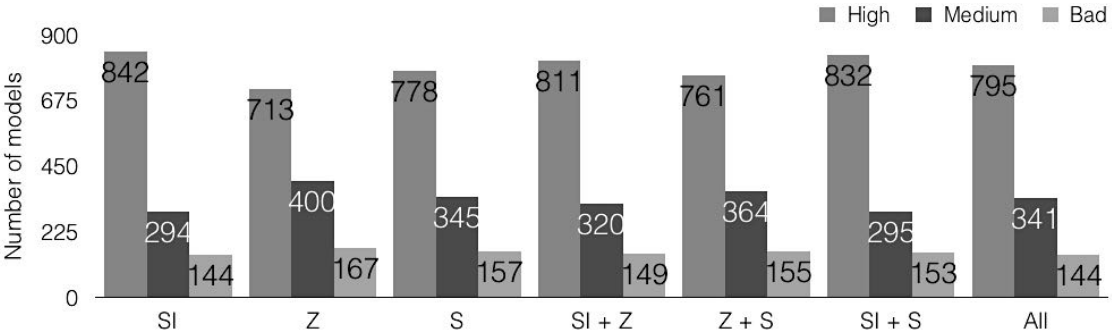


Figure 3.18: Quality of the top ranked models selected using different schemes. Sequence identity between query and template (SI), ZRANK global energy score (Z) and SOAP-PP protein interface statistical potential (S) were combined in different ways. The quality of the models were assessed by doing a structural comparison with the known structure. CAPRI criteria were used to classify the models into high, medium or acceptable based on $RMSD^{\text{interface}}$ and Fnat.

3.4.4.2 Comparison with Interactome3D

We compared the models produced by our template selection protocol to the models produced by Interactome3D. The Interactome3D protocol search for structural templates is based on the sequence identity returned by BLAST, rather than HMM based searches. Structural templates selected had a minimum global sequence identity of 40% with the query targets.

The qualities of Interactome3D models were obtained from the Nature Methods paper (Mosca et al., 2013a). However, since the benchmark target dataset used by Interactome3D is not publicly available, we compiled a benchmark from the November 2011 release of PDB (See Section 3.3.2.3) .

We compiled two benchmark datasets in a similar manner to Interactome3D based on (1) the Interactome3D November 2011 release (to compare with Interactome3D) and (2) the Interactome3D May 2015 release. We selected PPIs for which the query PPIs have been classified into CATH domains and a structural template could be found for both chains. Note that the size of benchmark target dataset built by the FunFam protocol (based on the November 2011) and Interactome3D is different because the FunFam protocol covers only protein complexes which comprised domains that can be mapped onto CATH-Gene3D. We did an analysis to determine the difficulty of the query targets found in the FunFam protocol derived dataset. A quarter of the query targets share sequence identity of less than 40% with the closest template. This demonstrates that the query dataset used to assess the FunFam protocol is more difficult (See Figure 3.19).

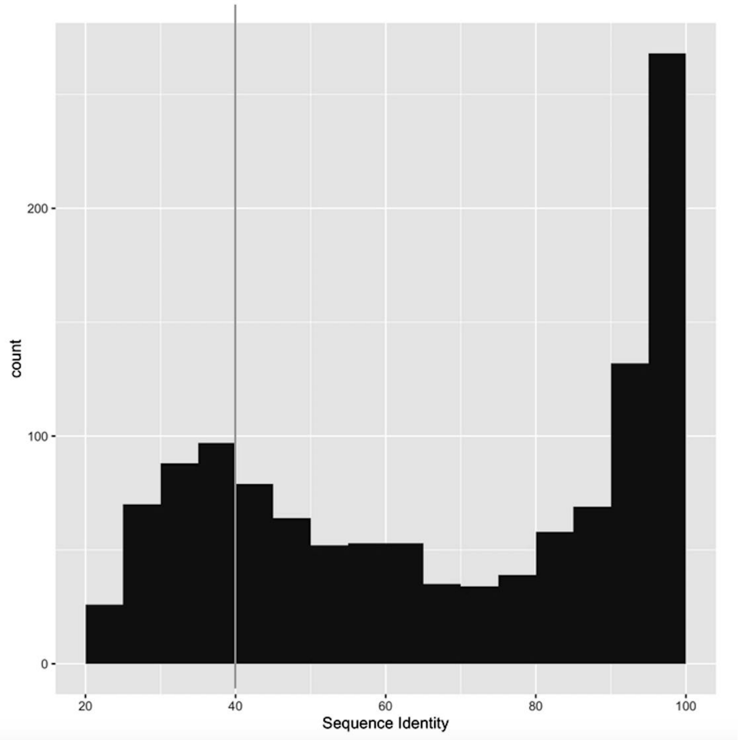
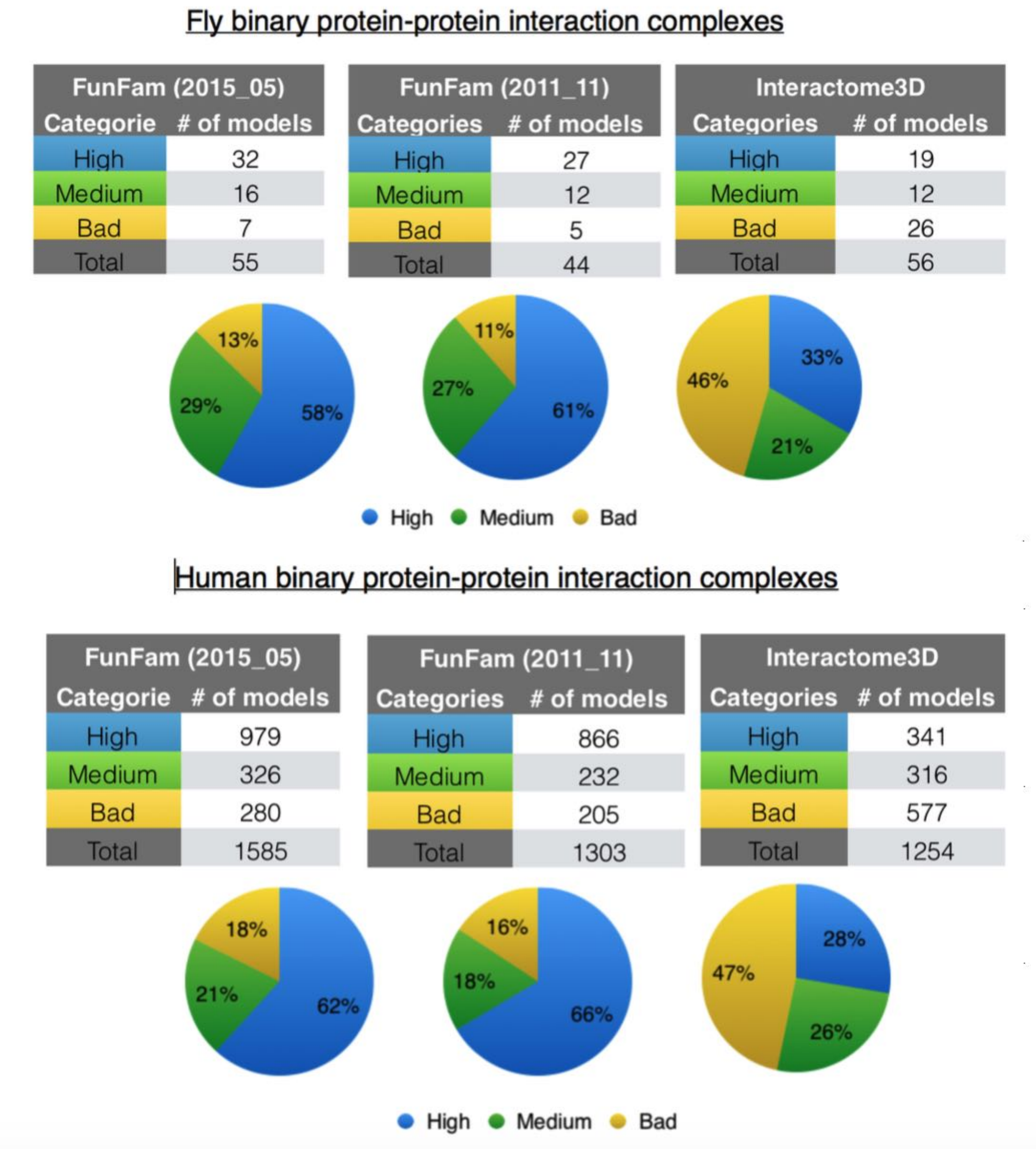


Figure 3.19: Difficulty of the benchmark target dataset built by the FunFam protocol. For the Interactome3D protocol, all targets in the benchmark set had more than 40% sequence identity with the templates.

Figure 3.20 shows a significant improvement in model quality using the FunFam protocol over the strategy used by Interactome3D. For the FunFam protocol 89% and 84% of the fly and human binary PPIs produced a medium- or high-quality model. In contrast, the top ranked models produced by Interactome3D gave medium- to high-quality model for only 55% and 52% of the fly and the human interactions. The FunFam protocol managed to produce 30% more medium/high quality models than Interactome3D. A higher proportion of the models produced by the FunFam protocol are of high quality. (See Figure 3.20) Clearly our HMM based protocol exploiting FunFams to select templates outperforms the simple BLAST based strategy of Interactome3D.



Fly binary protein-protein interaction complexes

FunFam (2015_05)		FunFam (2011_11)		Interactome3D	
Categorie	# of models	Categories	# of models	Categories	# of models
High	32	High	27	High	19
Medium	16	Medium	12	Medium	12
Bad	7	Bad	5	Bad	26
Total	55	Total	44	Total	56

Human binary protein-protein interaction complexes

FunFam (2015_05)		FunFam (2011_11)		Interactome3D	
Categorie	# of models	Categories	# of models	Categories	# of models
High	979	High	866	High	341
Medium	326	Medium	232	Medium	316
Bad	280	Bad	205	Bad	577
Total	1585	Total	1303	Total	1254

Figure 3.20: Comparison of the quality of the top ranked models produced by the modelling protocols of FunFams and Interactome3D. The quality of the models were assessed by doing a structural comparison with the known structure. CAPRI criteria were used to classify the models into high, medium or acceptable based on $RMSD^{\text{interface}}$ and Fnat.

3.4.4.3 Comparison with HHsearch

In Chapter 2, we demonstrated the advantages of using FunFams to search for structural templates of single domains, compared to HHsearch. In order to compare our

FunFam search strategy for templates of protein complexes with another powerful HMM based strategy, we used HHsearch to search for templates for the complexes. As before for both the FunFam and the HHsearch strategy we used MODELLER to build the models once we had selected the templates i.e. everything apart from the template selection strategy is the same.

For the FunFam pipeline, we first identified the FunFams that are most similar to the query, followed by the selection of the best template in the FunFams based on sequence identity and X-ray resolution. For the HHsearch pipeline, we used HHsearch to query for the best template and select the best template based on the program's built in statistical measures (E-value and Probability to assess the best template).

We used a different benchmark dataset from the previous Interactome3D analysis. We used the PBD5 dataset produced by Zlab. This benchmark dataset is designed for testing docking procedures and has been used widely by research groups to test protein complex modelling methods. We split all the chain-chain interactions in the benchmark using the CATH domain classification into 181 easy and 103 hard DDI targets.

The FunFam protocol could predict for 127 easy targets and 72 hard targets. HHsearch predicted 112 easy targets and 67 hard targets. Figure 3.21 demonstrates the overall percentage of good quality models built using the FunFam and the HHsearch protocol. For easy targets, HHsearch gave 6% more good quality models than FunFams. For the hard targets, FunFam gave 9% more good quality models than HHsearch.

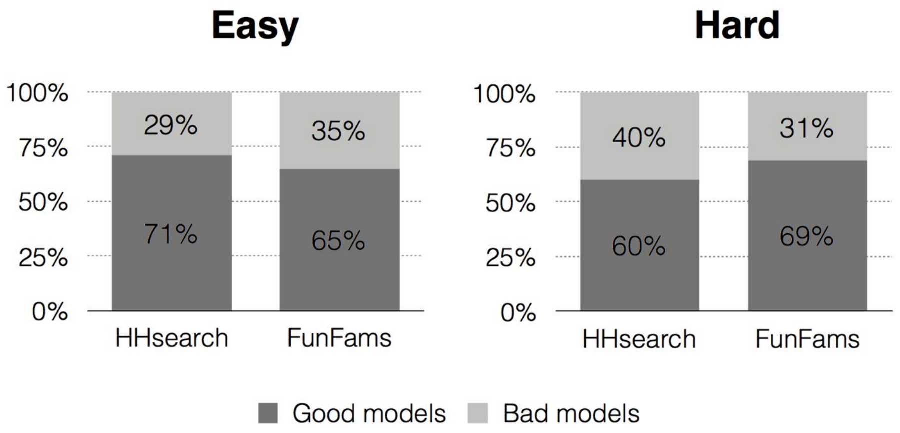
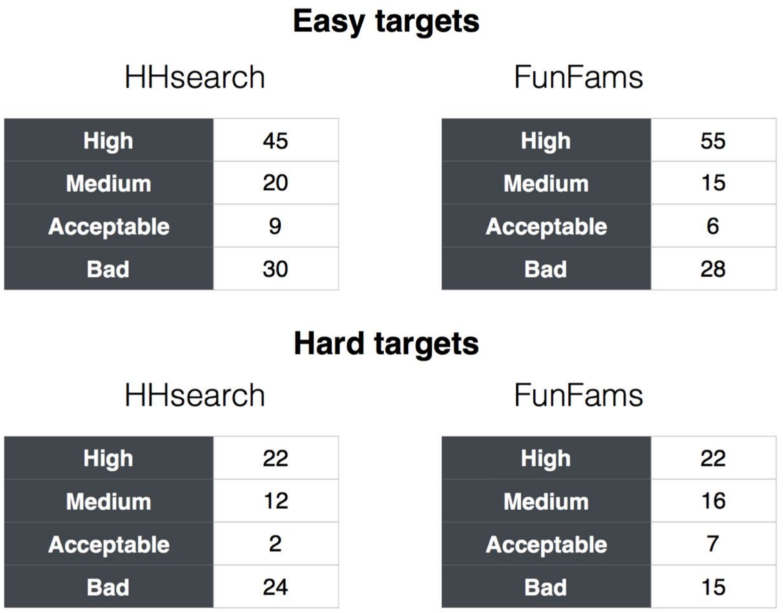


Figure 3.21: Quality of FunFams and HHseach models.

If we compared the common models built by FunFams and Hhsearch for both difficulty levels, FunFams models have a higher quality than the HHsearch models. However the differences are not statistically significant (Wilcoxon signed-ranked test, p-value $>$0.05).



Easy targets

HHsearch

High	45
Medium	20
Acceptable	9
Bad	30

FunFams

High	55
Medium	15
Acceptable	6
Bad	28

Hard targets

HHsearch

High	22
Medium	12
Acceptable	2
Bad	24

FunFams

High	22
Medium	16
Acceptable	7
Bad	15

Figure 3.22: Quality of common FunFams and HHsearch models

3.4.5 Modelling protein complexes using a 3D-Template based docking pipeline

In order to extend the number of protein complexes we could build i.e. human and fly, we developed a 3D-template based docking modelling pipeline adapted from the method established by the Vakser group (Kundrotas and Vakser, 2013). Vakser's method is one of the best 3D-Template based docking protocols at the moment. Since we did not have access to their template library, we decided to use DOCKGROUND template library. We first tested how well we implemented their method for use with the DOCKGROUND template library. In addition, we generated a new template library based on the FunFams. We also explored methods to rank the docking poses.

3.4.5.1 Benchmarking the template based docking pipeline

To compare against Vakser's protocol, we used their benchmark dataset of 372 non-redundant pairwise complexes. The main difference between our pipeline and Vakser's protocol is the template library used. We used the DOCKGROUND prebuilt publicly available template library of protein-protein interfaces (DOCKGROUND template library) in our study. This is the only publicly available resource of its kind.

Vakser group evaluated the models using the $RMSD^{\text{interface}}$ score calculated by superposing the modelled complex onto the known protein complex. The models were grouped into high-accuracy models ($RMSD^{\text{interface}}$<5Å) or low-accuracy models (5Å<$RMSD^{\text{interface}}$<10Å). Models with $RMSD^{\text{interface}}$ more than 10Å were considered bad.

The results obtained from the benchmark are shown in Figure 3.23. Using a similar target dataset, Vakser's template library gave more good models (i.e. high-accuracy + low-accuracy) than the DOCKGROUND template library (68% vs 57%). However, DOCKGROUND template library produced more high-accuracy models compared to Vakser's template library (40.2% vs 38.4%). Note that since the Vakser library is not publicly available, the results in Figure 3.23 for the Vakser library were taken from a publication by Vakser group of results obtained using the same target dataset.

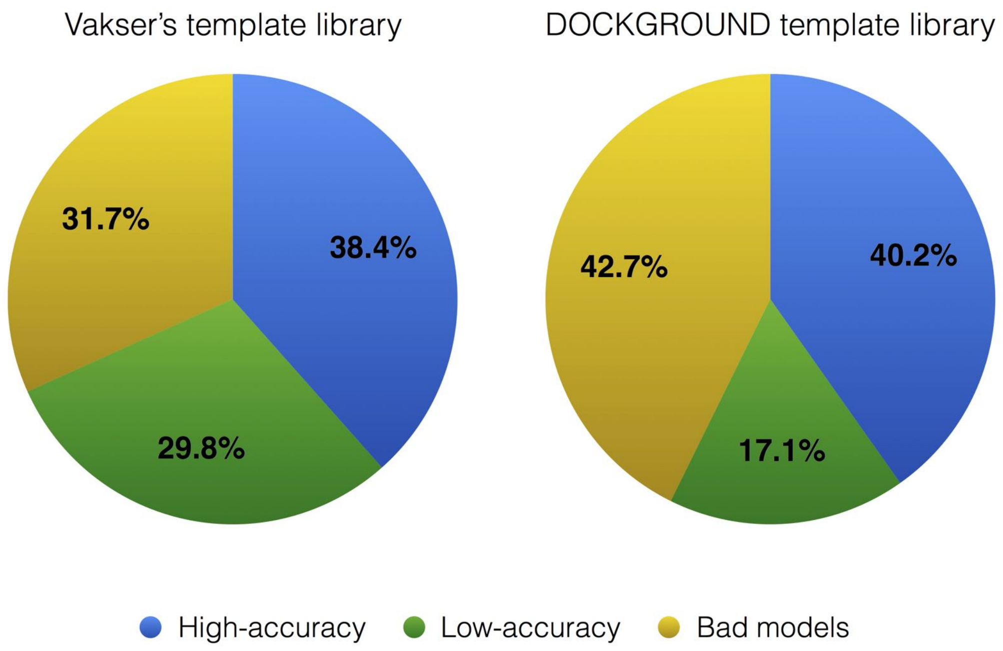


Figure 3.23: Quality of models built using the DOCKGOUND template library compared to Vakser's template library

Our results showed that we had implemented the Vakser protocol well and obtained comparable results. Next, we endeavoured to improve the performance of our pipeline by using a better template library. The DOCKGROUND template library has not been updated since July 2012. Therefore, we decided to generate an in-house template library dataset. Generating an in-house self-generated template library gave us more control over the data. For example, different structural redundancy criteria could be used, or possibly NMR solved protein structures could be included.

3.4.5.2 Benchmarking the Lam Template Library

The Lam library has a total of 12,240 pairwise complexes. The major differences between the Lam library and the DOCKGROUND library are the PDB dataset used, the clustering method of the protein complexes and the removal of illegitimate complexes, as detailed below. DOCKGROUND library used the July 2012 PDB release, whereas the Lam library uses the July 2014 PDB release entries. We speculated that using a newer PDB dataset would include additional structures which could improve the quality of models built, as closer homologues may be available as templates.

DOCKGROUND used a structural similarity threshold of TM-score of 0.9 to reduce the redundancy of the whole dataset. In building the Lam library, FunFams were employed to cluster the protein complexes. However, unlike the DOCKGROUND library, the Lam library does not remove illegitimate complexes such as interwoven chains, tangled chains and chains with interacting unfolded termini parts. This is due to the difficulty of extracting the information from the PDB file.

Table 3.9: Differences between template datasets

Template dataset	Number of interfaces	Clustering criteria
DOCKGROUND template library	7,107	Structural similarity
Lam template library	12,240	Structural and functional similarity (FunFams)

In the previous section, we used a benchmark target dataset compiled by the Vakser group because we wanted to assess the performance of the Vakser method using a DOCKGROUND library and ensure we had implemented the Vakser method correctly. However, this target dataset is not used extensively by groups researching docking other than the Vakser group. Therefore, we employed a new benchmark which uses the docking target dataset PBD4 (Hwang et al., 2010) that has been used extensively by other research groups.

Figure 3.24 demonstrates the quality of models built using the two libraries. The models were assessed using the CAPRI criteria given in Section 3.3.4. The Lam's template library gave 37% high, 11% medium, and 4% acceptable quality models. The Lam's library gave a total of 57 good models compared to 45 for the DOCKGROUND library.

We believe that the improvement of model quality may be due to the usage of FunFams in removing the structural redundancy of the protein complexes in the template library. By using the structurally and functionally coherent FunFams, the Lam library captured more diversity of pairwise complexes than the DOCKGROUND template library which was clustered solely on structural information.

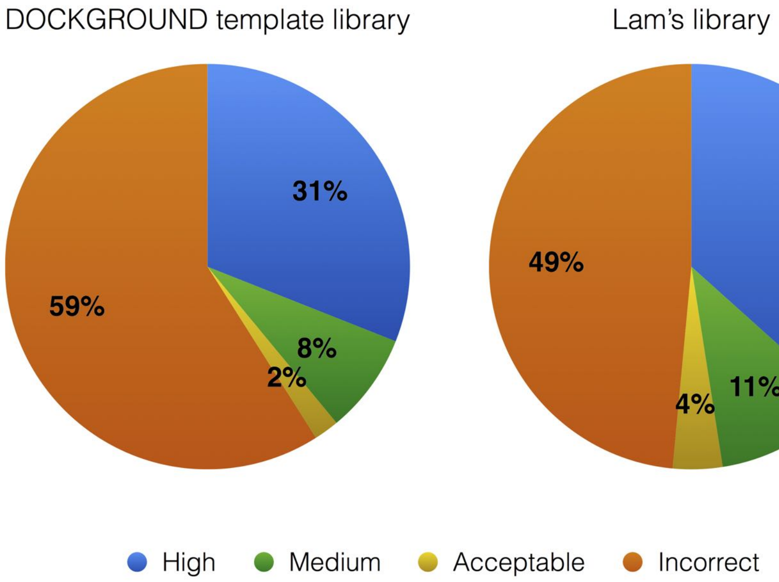


Figure 3.24: Proportion of targets that gave models of different quality built by different template libraries

3.4.5.3 Ranking the docking poses

Unlike the modelling of individual proteins, where the practice of selecting templates is to select the closest sequence relative with a known structure, when modelling complexes the best criteria for selecting templates is less well understood (Kastritis and Bonvin, 2010; Vreven et al., 2012).

Therefore, the usual practice, adopted by Vakser and others, is to select all templates that meet the match criteria, build the models based on these templates and then rank the models based on how well they capture known features of complexes (e.g. global energy value or structural similarity, see Table 3.5).

Therefore we evaluated different ways of ranking the models as this will affect the performance of the pipeline since good models (i.e. superposing well on the native structure) should be ranked highly. Certainly it is important that at least one acceptable model (Fnat $\geq$10% and 2.0Å$<RMSD^{\text{interface}}\leq$4.0Å) is in the top 10 ranked models for a target.

The models were ranked using sum of TM-scores (SoTM) according to the decreasing order of the sum of the two TM-scores (one for each target monomer) obtained from the alignment between docking targets and docking templates (Negroni et al., 2014).

Earlier analysis (see Figure 3.24) showed that 57 targets out of the PBD4 target dataset produced at least one acceptable model using the Lam template library. Therefore, the following analysis will focus only on these targets. Using SOTM, only 20% of the targets had at least one good model in the top 10. This is possibly due to the structural redundancy in the template library. Therefore, different filtering strategies were tried to solve this issue.

MODELLER's normalised DOPE (nDOPE) was used to identify and remove any bad models. Models with a postive nDOPE score are likely to be poor, while models with scores lower than -1 are likely to be native-like (Eswar et al., 2008). Therefore, all models that have nDOPE score greater than 0 were removed. Using this filtering the results improved with 35% of the targets now having at least one acceptable model in the top 10.

Subsequently, models that possess severe steric clashes were also eliminated. They are models with more than 3 interchain Cα-Cα within a distance of 3Å (Negroni et al., 2014). We also removed structural redundancy in the models. Pairwise structural comparisons between models were done using the MM-align method (Mukherjee and Zhang, 2009). MM-align was developed by the Zhang group to structurally compare two multi-chain protein complexes. We decided to remove redundant models that score TM-align more than 0.95. CATH superfamily information was also used to remove redundant models. The following table summarizes the results obtained using the various filtering criteria.

Table 3.10: Number of targets with at least one good model in Top 1, 5, 10 after applying different criteria

Methods	Top 1	Top 5	Top 10
SoTM	0 (0%)	5 (9%)	11 (20%)
SoTM + nDOPE	5 (9%)	15 (27%)	19 (25%)
SoTM + nDOPE + clashes	9 (16%)	21 (38%)	22 (40%)
SoTM + nDOPE + clashes + MM-align	9 (16%)	22 (40%)	23 (42%)
SoTM + nDOPE + clashes + MM-align + CATH	9 (16%)	22 (40%)	24 (44%)

3.5 Conclusions

Since the structural coverage of known protein-protein complexes is relatively low, structural modelling of protein complexes is important. Structural annotation of protein complexes gives us a better understanding of the composition and ultimately the function and mechanism of the protein complexes.

We began by exploring the structural conservation of protein-protein interfaces in FunFams. Our results demonstrated that within FunFams which are clusters of functionally an structurally coherent relatives, domain-domain interfaces are more conserved than across a superfamily. The difference between conservation within FunFams and across a superfamily is particularly pronounced for inter-chain interfaces especially at low sequence identity. However, there were problems for some of the FunFams, and we speculated that the low structural conservation for some DDIs are due to pollution of the FunFams by relatives having different MDAs, and other possible issues such as different ligands bound.

Next, we explored the usage of FunFams to model protein complexes using MODELLER (sequence based methods). FunFams were used to identify the structural template. We managed to produce 30% more medium- to high-quality models for fly and human PPIs than Interactome3D which uses BLAST to select templates. In addition, a higher proportion of the models produced by our protocol are of higher quality than Interactome3D. The FunFam protocol built models are slightly better than HHsearch built models but this was not statistically significant. However, our results demonstrated that FunFams can facilitate the template selection process. In addition, by using FunFams we can use the MSAs of relatives in the FunFam to provide information on conserved surfaces which could help validate the prediction of the interface.

We also devised a new template library of pairwise complexes for template-based docking using FunFams. The template library gave 10% more good quality models than the widely used publicly available DOCKGROUND template library.

3.5.1 Research outlook and recent developments in protein complex modelling

We demonstrated the rise of co-evolution in guiding protein monomer structure modelling in Chapter 2. Similarly, co-evolution information can also be utilised for complex modelling. Guerois group (Yu et al., 2017), the best non-server structural modelling group in the recent CAPRI rounds (Lensink et al., 2017), use co-evolution information in their protocol. Their new statistical potential employed for scoring docking poses is complemented with co-evolution information. Also, the docking poses were filtered using different scoring functions that included a co-evolution score (InterEvScore) (Andreani et al., 2013) and a statistically optimized SOAP potential (Dong et al., 2013).

Analyses of protein-protein interfaces show that native conformation do not necessarily have the most hydrogren bonds or the largest buried solvent accesible surface area (Im et al., 2016; Norel et al., 1999; Xu et al., 1997). Some proteins involve changes in the shape to enable interactions (Induced fit hypothesis). The Zacharias group developed a coarse-grained force field that has the ability to incorporate conformational flexibility during the initial sampling step of docking poses and also an energy minimization step (Schindler et al., 2017). Other research groups are likely to employ this type of information in the future.

Template based docking is effective when dealing with protein complexes having sufficiently close homologues with known structure. However, if no suitable homologues can be found, template-free docking needs to be employed. The Huang group developed hybrid methods that combine both template-based and template-free methods (Yan et al., 2017). Incorporating template based docking in the protocol decreases the amount of sampling space. If the template is reliable, template free docking will be able to tweak the models into a better binding mode. Homologous complexes do not guarantee similar binding interfaces or binding orientations. If the template is wrong, template free docking can still sample correct binding modes and produce a correct prediction. In contrary, Seok group added *ab initio* refinement steps to refine the results of template based docking (Lee et al., 2017).

Further investigation should be carried out to improve the ranking process. ZRANK2 can be used to remove bad models based on physics based potentials such as desolvation energy, Van der Waals, partials electrostatics (Pierce and Weng, 2007). Models with low global energy can be removed using FiberDOCK (Mashiach et al., 2010). Recently, there are also modelling groups that use machine learning methods to rank docked protein complexes. The Bates group developed a clustering method that utilised a randomised tree classfier based on 109 molecular distributors and obtained good results in the recent CAPRI rounds (Lensink et al., 2017; Pfeiffenberger et al., 2017).

Although there has been significant improvement over the years in complex modelling, there are still very challenging cases that involve multiple binding patches, cases that involve small interface regions or cases where interface residues lie in loop regions (Lensink et al., 2017; Soni and Madhusudhan, 2017). Further research is clearly still needed to improve the complex modelling.

Chapter 4

Protein structure and function analyses to understand the implications of alternative splicing

4.1 Introduction

Alternative splicing (AS) has been widely recognised as one of the major processes expanding the diversity of proteomes in multicellular organisms. AS refers to the assemblage and rearrangement of different exons (expressed sequences, which code for proteins) and introns (intervening sequences) of a gene during the pre-mRNA splicing such that different mRNAs and thus, proteins are produced from the same gene. It occurs predominantly in higher eukaryotes (producing different isoforms in different developmental stages, tissues or disease states) while keeping the number of genes relatively low (Pohl et al., 2013).

AS is usually mediated by a large molecular machinery named the spliceosome. The spliceosome recognises the exons and introns by using three major sequence elements: the 5' donor site, the 3' acceptor site and a branch point. The whole splicing process involves two sequential transesterification reactions (the process of exchanging the organic R group of an ester with the organic R group of an alcohol) between mRNA nucleotides. First, the 2'OH of the branch point performs a nucleophilic attack on the first nucleotide of the intron at the 5' donor site, forming a lariat. Next, the 3'OH of the release 5' exon performs the second nucleophilic attack at the first nucleotide following the last nucleotide of the intron at the 3' acceptor site, joining the exons and releasing the intron lariat. (See Figure 4.1)

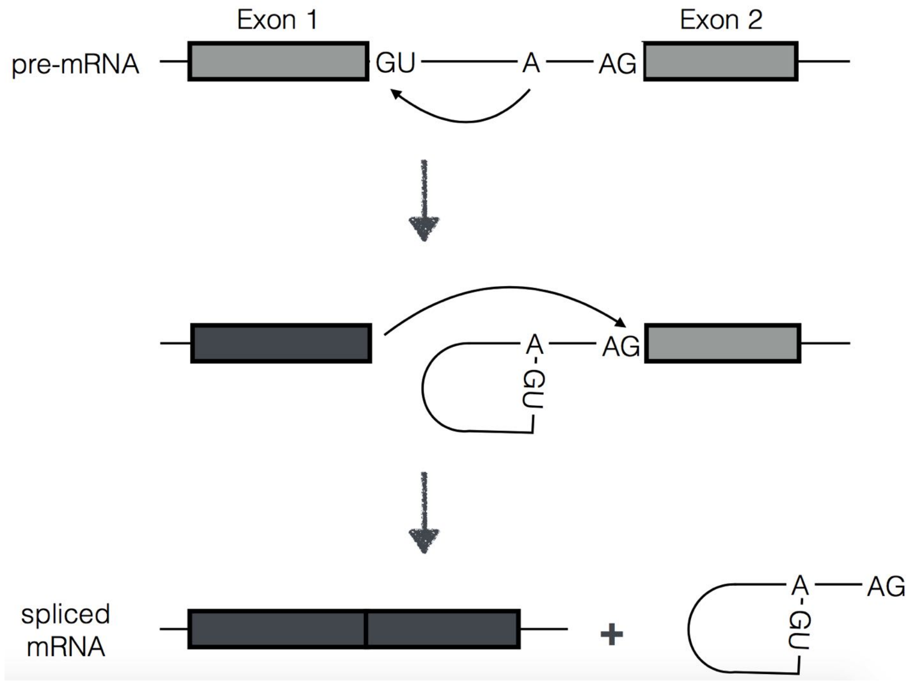


Figure 4.1: The two stages of splicing. In the first step of splicing, the 2' OH of an intronic adenosine of the branchpoint attacks the phosphodiester bond at the 5' donor site to generate a free exon and intron-exon intermediates. At this point, the intron is in the form of a lariat containing a 2'-5' bond between the 5' end of the intron and the branchpoint. In the second step of the reaction, the 3' OH of free exon attacks the phosphodiester bond at the 3' splice site to yield the ligated exons and lariat intron.

Figure 4.2 illustrates the five basic types of AS events. Exon skipping is a form of AS where the cells "skip" the exon region/regions. This is the most common mode of AS and accounts for nearly 40% of AS events in higher eukaryotes (Alekseyenko et al., 2007; Kim et al., 2007; Sugnet et al., 2004). Alternative donor site AS happens when an alternative 5' donor site is used and changes the 3' boundary of the upstream exon. This type of splicing event accounts for about 8% of AS events in higher eukaryotes (Sugnet et al., 2004). By contrast, alternative acceptor site AS uses an alternative 3' acceptor site and accounts for about 18% of AS events (Sugnet et al., 2004).

Intron retention AS usually happens when the mRNA is not flanked by introns. This confuses the splicing machinery, therefore a part of the sequence may be spliced out as an intron or retained as mRNA. Mutually exclusive exons are characterised by the splicing of exons in a coordinated manner such that two or more splicing events are not independent. Only one out of the two exons is retained, while the other one is spliced out. Both intron retention and mutually exclusive exons are classified as the less frequent AS events (Keren et al., 2010; Sugnet et al., 2004).

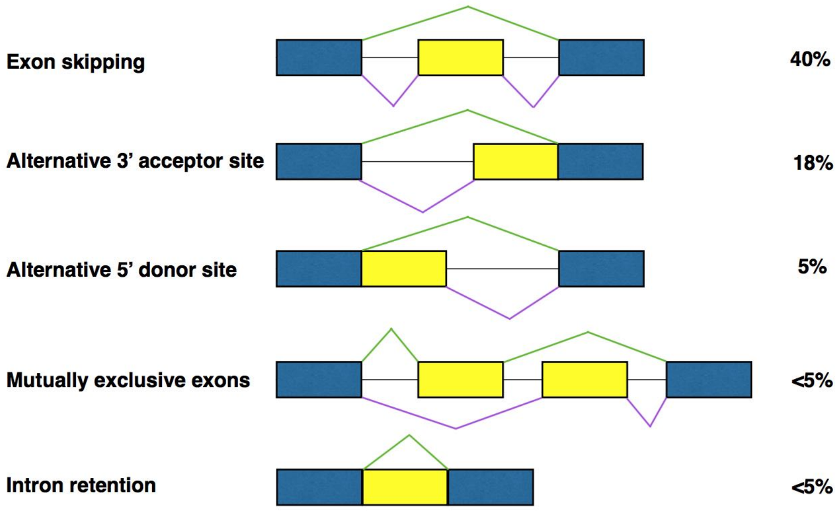


Figure 4.2: Different modes of AS and their relative abundance in higher eukaryotes. Exon skipping "skip" the exon region/regions. Through the use of alternative 5' or 3' splice sites, exons can be extended or shortened in length. One exon or the other is included, but not both in mutually exclusively exons. Finally, the excision of an intron can be suppressed, to leave the retained intronic sequence in the mRNA that is exported to the cytoplasm. Constitutive exons are shown in blue and alternatively spliced regions in yellow, introns are represented by solid black lines, and solid green/purple lines indicate splicing activities.

The presence of alternatively spliced transcripts has been proved experimentally by various techniques such as microarray and RNA-Seq studies (Graveley et al., 2011; Sánchez-Pla et al., 2012; Uhlén et al., 2015). A recent large-scale RNA-seq experiment demonstrated the existence of different splice variants for 72% of the annotated human genes (Uhlén et al., 2015). Another research that integrated *ab initio* gene prediction and RNA-seq identified over 205,000 human transcripts with protein-coding potentials (Hu et al., 2015). For example,the Drosophila Down Syndrome Cell Adhesion Molecule gene (DSCAM) has 4 variable exon clusters with one of them involving 48 alternative exons. This leads to 38,016 potential splicing isoforms (Schmucker et al., 2000; Wang et al., 2012) and allows *D. melanogaster* to exhibit a unique set of DSCAM proteins on its cell surface that are essential for immunity and normal neuronal processes. The overpression of the human homolog of DSCAM in developing fetal neural tissue can leads to Down syndrome (Schmucker and Chen, 2009; Yamakawa et al., 1998).

Proteins produced by protein isoforms may have modified structures and functions. Tress et al. demonstrated that AS can lead to wide range of structural outcomes, and most of them are undesirable, and potentially deleterious (See Figure 4.3 for examples of potential effects). It was found that up 60% of the AS isoforms that could be modelled by Tress and co-workers, lacked a major part of a domain (2007). Light and Elofsson showed that 36% of the isoforms in Swiss-Prot alter the domain architecture of the proteins (gain/loss of a globular domain, transmembrane anchor).

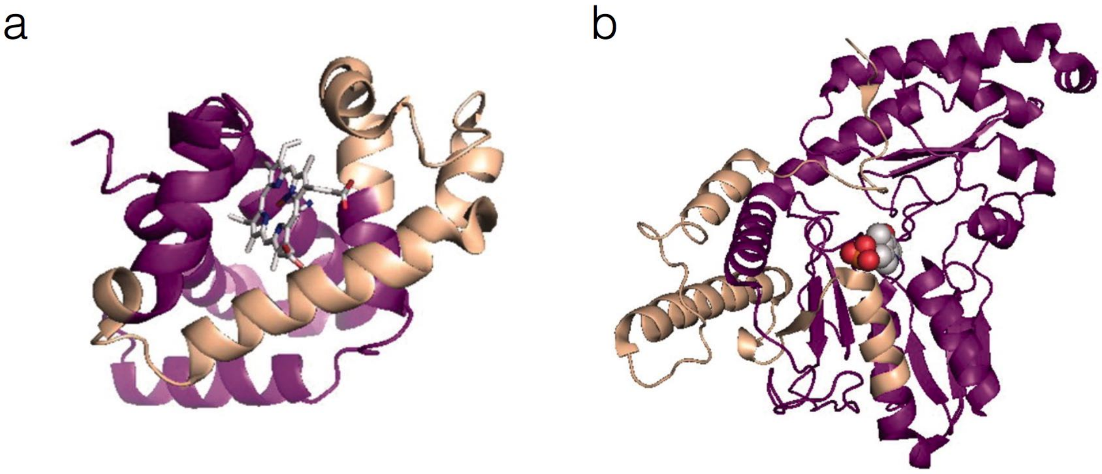
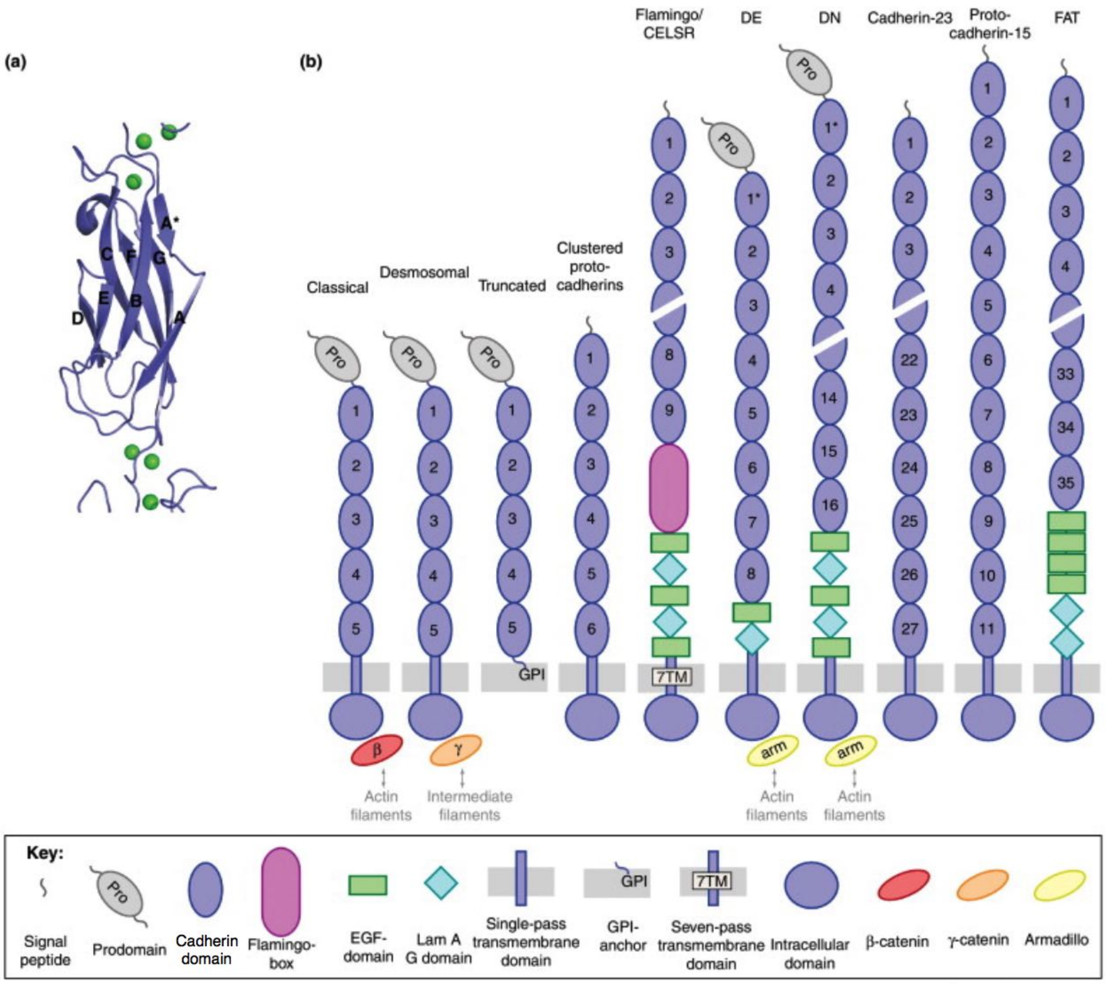


Figure 4.3: Potential effects of AS on protein structure. Splice isofroms were mapped to the closest structural templates. Structures are colored in purple where the splice isoform is missing. (a) Haemoglobin δ-subunit isoform mapped onto PDB structure 1si4. (b) Mitochondrial cysteine desulfurase isoform mapped onto PDB structure 1p3w. (taken from Tress et al. 2007).

Liu and Altman identified a total of 24 protein domains that more commonly undergo AS than other proteins in human (2003). The most prevalent domains were the repeating cadherin domain and domains that involved intrinsically disordered regions (Liu and Altman, 2003; Tress et al., 2008). Cadherin (named for "calcium-dependent adhesion") is a type of transmembrane protein which is predominantly involved in the process of cell communication/signalling, development and apoptosis. Figure 4.4 illustrates the members of the cadherin familiy which share the same cadherin domain.



TRENDS in Cell Biology

Figure 4.4: Schematic representation of members of the cadherin family, which share the common cadherin domain. (a) Typical cadherin domain (b) Schematic representation of the domain organisation of selected cadherin family members (taken from Brasch et al. 2012).

Intrinsically disordered regions are prominently found among the hubs in a PPI network and it is believed that changes in these proteins are important for mediating the remodelling of PPI networks (Buljan et al., 2012; Ellis et al., 2012). For example, protein phosphatidylinositol-4-phosphate 5 kinase, type 1, gamma (PIP5K1C) is spliced differently in cerebellum and lymph node. In the cerebellum, a tissue-specific disordered region is spliced into the protein. This region acts as the binding site for adaptor protein complex 2 (AP-2) (Thieman et al., 2009). This initiates further interactions between AP-2 and other proteins, that are important for the process of synaptic vesicle transport. By contrast, the tissue-specific segment does not get spliced into the PIP5K1C protein in lymph nodes. Therefore, PIP5K1C is unable to interact with the AP-2 complex and the whole interaction network is thus rewired. PIP5K1C, in this case, is involved in cell migration (Buljan et al., 2012). Figure 4.5 illustrates this rewiring process.

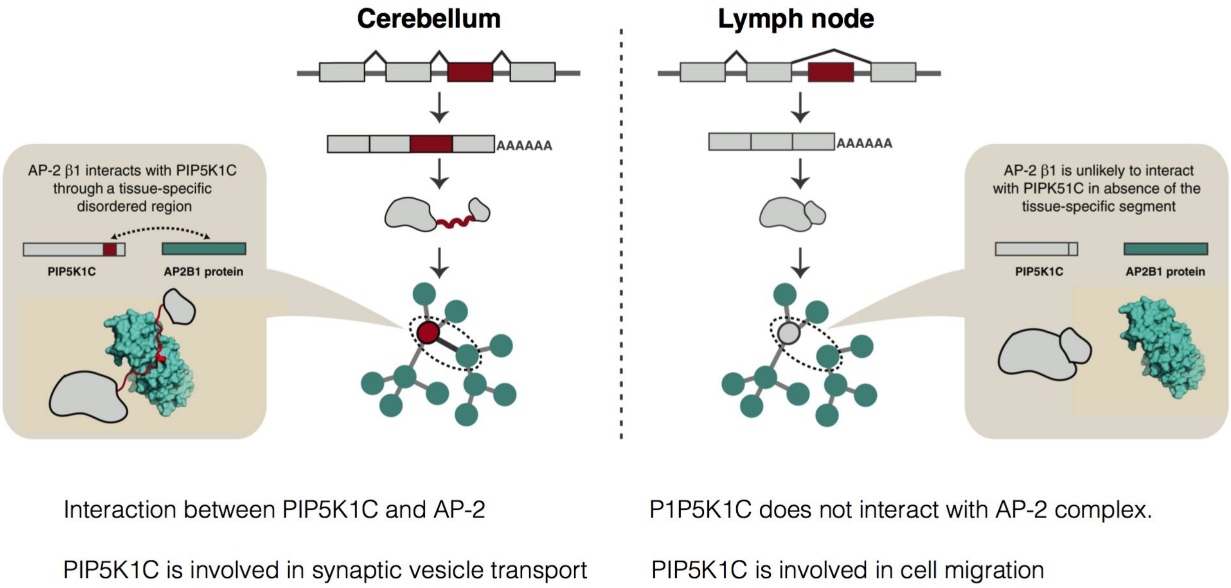

Cerebellum
Lymph node
AAAAAA
AAAAAA
AP-2 β1 interacts with PIP5K1C through a tissue-specific disordered region
PIP5K1C
AP2B1 protein
AP-2 β1 is unlikely to interact with PIPK51C in absence of the tissue-specific segment
PIP5K1C
AP2B1 protein
Interaction between PIP5K1C and AP-2
PIP5K1C is involved in synaptic vesicle transport
P1P5K1C does not interact with AP-2 complex.
PIP5K1C is involved in cell migration

Figure 4.5: Tissue-specific splicing of a disordered segment rewires the protein interaction network. This is illustrated with the PIP5K1C kinase gene that has an exon with different inclusion levels in cerebellum and lymph mode. The exon encodes a binding motif (red cartoon representation), which mediates the interaction with the AP-2β appendage domain (cyan surface representation). (adapted from Buljan et al. 2012).

AS has been linked to human diseases and cancer (Dutertre et al., 2010; Mayr et al., 2011; Tang et al., 2011). An example of AS-related diseases is CRASH (Clinical spectrum of corpus callosum hypoplasia, Retardation, Adducted thumbs, Spastic paraparesis and Hydrocephalus) syndrome, attributed to the splice isoform 6 of L1 cell adhesion molecule (L1CAM) (Gabellini et al., 2006). L1CAM is involved in neuronal adhesion, neurite outgrowth, and axon guidance. This transcript does not include exon 2 (an extracellular domain, which encodes the sequence YEGHHV, located immediately N-terminal to the first Ig domain) into the L1CAM protein. The loss of this exon 2 means that the alternative isoform exhibit reduced binding of cell adhesion molecules and reduced neurite growth which leads to the manifestation of the syndrome (Jacob et al., 2002).

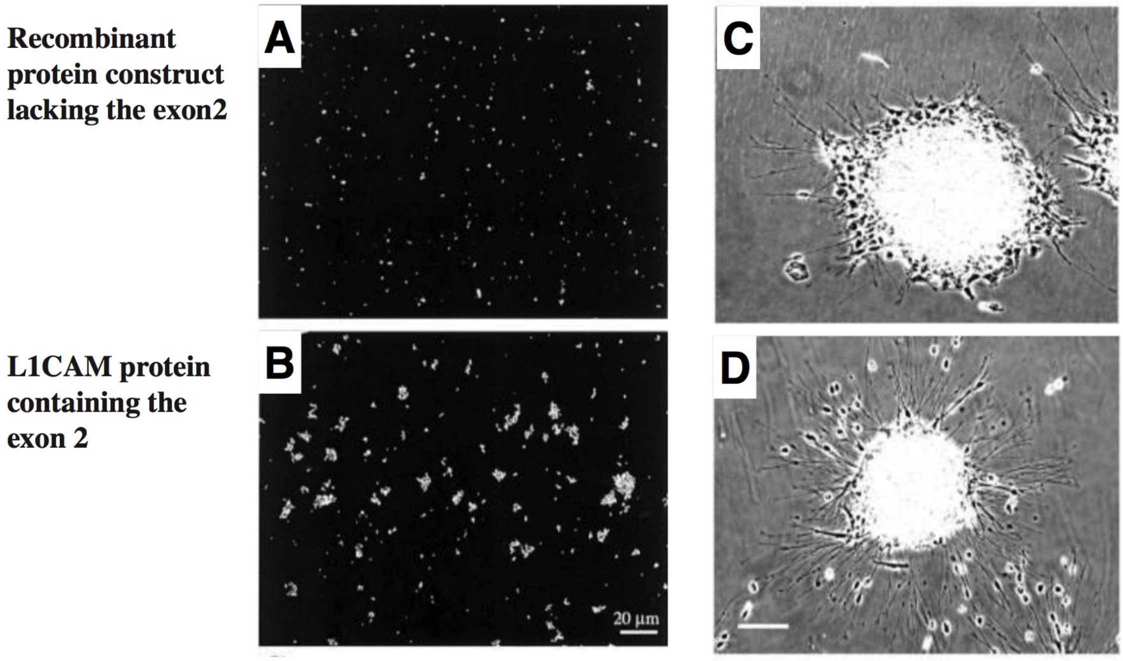

Recombinant protein construct lacking the exon2
L1CAM protein containing the exon 2
A
B
C
D
20 µm

Figure 4.6: When bound to fluorescent beads, recombinant protein lacking exon2 (A) exhibits less binding compared with isoform which has exon2 (B). The neurite outgrowth is also diminished in cells lacking exon2 (C) compared to cells in which the L1CAM protein has exon2 (D) (taken from Jacob et al. 2002)

Another example is Collagen alpha-3 type VI protein (COL6A3). COL6A3 is involved in cell anchoring and remodelling of the extracellular matrix (Zanussi et al., 1992). Studies have shown a higher level of the COL6A3 proteins in tumor tissues than normal tissues (Arafat et al., 2011; Nanda et al., 2004; Xie et al., 2014). Recent genome exon array studies have identified tumor-specific exons 3, 4, and 6 of COL6A3 in colon cancer, and exon 6 in colon, bladder, and prostate cancers (Gardina et al., 2006; Thorsen et al., 2008). The cancer causing exon 6 of COL6A3 contains a von Willebrand factor domain with seven predicted phosphorylation sites, which have a potential regulatory effect on the protein function (Thorsen et al., 2008). There was also a study that showed a significant upregulation of COL6A3 (exons 3 and 6) in human biopsies and cell lines derived from pancreatic cancer (Arafat et al., 2011). These suggest a role for COL6A3 in tumor formation.

High-throughput tandem mass spectrometry (MS) based proteomics has been the main tool for validating the evidence of AS at the protein level. Some groups have shown that MS only verifies a small number of alternative protein isoforms in different species (Brosch et al., 2011; Tanner et al., 2007; Tress et al., 2008). In a recent large-scale proteomic analysis, the Valencia group identified splice isoforms for 150 out of 22,304 human genes (Ezkurdia et al., 2012). In contrast, some other proteomics studies have reported substantially more cases of AS at the protein level. For example, Wilhelm and co-workers identified 1,297 alternative proteins for more than 18,097 human genes (7%) (2014), and, Kim and co-workers found 2,861 protein isoforms from more than 17,294 human genes (16%) (2014). To account for these contradictory findings, the Valencia group carried out a reanalysis of the peptides from eight large-scale MS experiments. They used a stringent filter to remove false-positive peptides by removing all the peptides that were only identified in one study. They also removed peptides that mapped to more than one gene. In the end, they recovered a total of 246 genes (282 events) that have reliable evidence for more than one isoform, for 12,716 protein-coding genes (Abascal et al., 2015a). This result suggests that alternative variants are not abundant at the protein level.

However, there may be some technical issues present in these current MS experiments. A typical MS experiment can only identify a small proportion of the peptide ions present in a protease digest. This means that MS will fail to detect all proteins that are expressed especially those that are expressed in low quantities (Abascal et al., 2015a). This may explain the low number of alternative isoforms in proteomics experiments. The other technical issue is the limited sampling depth (Abascal et al., 2015a). Some splice isoforms are only expressed in limited tissues, or under special circumstances. Another possible reason is that some of the splice isoforms may have other functions than generating protein product (such as nonsense-mediated mRNA decay) (Lareau and Brenner, 2015).

Although the splice isoforms identified in proteomics experiments have low coverage, they have been found to be highly enriched with mutually exclusively (MXE) exons both in human (60 out of 282 splicing events) and mouse (21 out of the 68 splicing events) analyses (Abascal et al., 2015a). It is usually believed that MXEs originated from exon duplication, and often show high sequence similarity. They are usually similar in the length (typically fixed for α-helical and β-strands regions, but flexible in loop regions), conserved splice site patterns (e.g. GT-AG, GC-AG, and AT-AC), same reading frame and sequence homology (Hatje and Kollmar, 2014; Pillmann et al., 2011). Based on these assumptions, computational methods have been developed to predict MXE events (Pillmann et al., 2011; Stephan et al., 2007). Kassiopeia is a repository composed of predicted MXE events for 12 *Drosophila* species, *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Homo sapiens* with stringent parameters (Hatje and Kollmar, 2014).

The Valencia group performed an analysis of MXE events to measure the structural and functional effects of AS by analysing the composition of conserved Pfam functional domains in the predicted protein products. Only 1 in the 60 human MXE events broke a Pfam domain (measured by loss or gain five or more residues) (Abascal et al., 2015a). They managed to map 9 of the 60 human MXE to known PDB structures. These MXE splice events are usually believed to have subtle effects on the protein folds. They may affect the binding of the protein to another substrate (protein, ion) (Abascal et al., 2015a) or disturb the catalytic site or an allosteric site (See Figure 4.7 for an example where MXE region lies close to an functional site). MXEs have been proven to be of biological relevance (regulation of mammalian pyruvate kinase expression level (Chen et al., 2012), the voltage dependence of ion channels (Soom et al., 2008)).

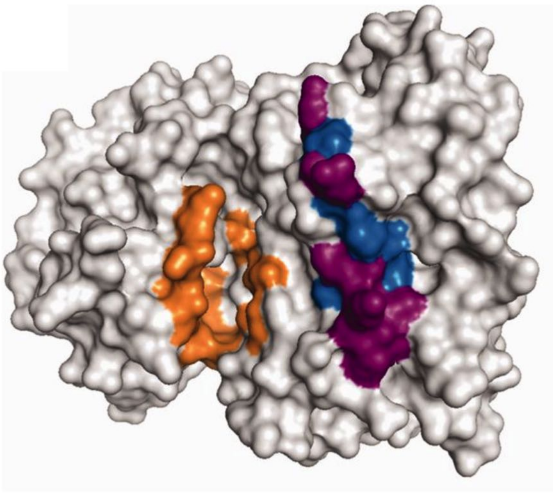


Figure 4.7: An example showing an MXE region, in which the amino acid residue usage differs between alternative isoforms (blue and purple), that lie close to the location of an ligand-binding site (orange) in human MAPK8 (taken from (Abascal et al., 2015b)). The biological significance of this MXE may relate to different ligand-binding specificities (Gupta et al., 1996).

4.2 Objectives

The main aim of the analysis was to understand the implications of AS on human and fly protein structures. We were particularly interested in the MXE splice variants, since these types of events are more supported by experimental analysis and proteomics data than other types of AS. Furthermore, most MXE events do not involve big structural changes of the protein fold, but changes in residue usage. For the fly dataset, we downloaded all the MXE annotations from the Kassiopeia resource and translated gene sequences into amino acid sequences. For the human dataset, we identified MXEs using an in-house method (See Methods section). We then assigned the different splice isoforms to CATH FunFams and annotated these isoforms with structural information. We determined whether the variable residues between exon pairs are exposed to the solvent by calculating the relative solvent accessibility surface areas of these residues. We also analysed if these variable residues are in the vicinity of any known functional sites such as catalytic residues (from the Catalytic Site Atlas(CSA)), PPI sites and protein-small molecule interaction sites (from the Inferred Biological Interactions Server(IBIS)), allosteric sites (predicted using the

A-SITE predictor, developed by Dr. Aurelio Moya Garcia in our group) and FunSites (our in-house functional site predictor, proven to predict sites enriched with CSA and IBIS PPI residues). We then used the McLachlan physiochemical matrix to quantify the amino acid changes of variable residues. We also mapped residues mutated in cancer, from the COSMIC database to MXE events to determine if these genetic variations lie in the vicinity of MXE events.

4.3 Materials and Methods

4.3.1 Identifying the amino acid sequences for MXE events

4.3.1.1 Fly

We downloaded the annotations of predicted MXEs for *drosophila melanogaster* from the Kassiopeia resource which corresponded to Flybase 5.36 and kept the MXE events and assoicated protein sequences that we could map to this genome. By referring to FlyBase version 2017_02, gene sequences and coordinates (32), we extracted all the corresponding amino acid sequences. For example, we only use predicted exons in the core genome annotation and only used examples where the translations from chromosome sequence exactly matched the Flybase translation. Further annotations of the MXE events (FlyBase Gene id, FlyBase protein id, FlyBase transcript id) were added using the ENSEMBL BioMarts tool (Kinsella et al., 2011).

4.3.1.2 Human

The Kassiopeia resource does not provide MXE annotations for human genes. In order to identify human MXEs, we did the following. Based on Ensembl version 87, we obtained the human mRNA transcripts (from the General Transfer Format) and identified sets of human genes with MXEs. We selected the longest transcript as our reference and compared all the other transcripts against the reference to look for MXE exons (See Figure 4.8). As mentioned already MXE have been shown to be highly similar in sequence and length. Therefore, we made sure that the exons were consecutive in the DNA sequence, but never occurred together in the same transcript, were similar in length and had the same reading frame (See Figure 4.8). After that, we translated all the MXE exons into amino acid sequences and compared the sequences against each other with BLASTp, setting an e-value threshold of 0.005 and sequence identity 25%.

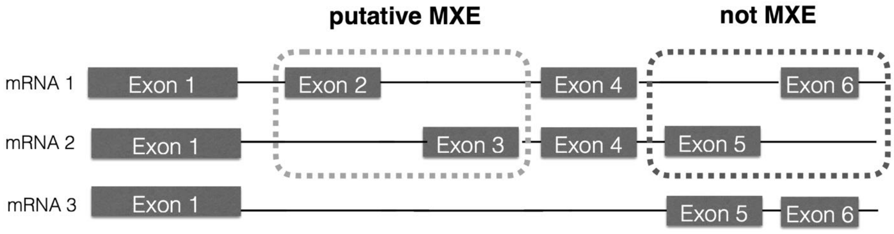


Figure 4.8: Identifying the putative MXE exons. We compared all the transcripts and made sure the MXE exons identified are never co-occur sequentially in any transcript, similar in length and had the same reading frame and sequence homology.

4.3.2 Predicting the function of MXE genes

We used the version 6.8 of DAVID (database for annotation, visualization and integrated discovery) function annotation tool suite (Huang et al., 2009) to identify functional enrichments of identified MXE genes. DAVID is widely used by the scientific community with over 21,000 citations. The program first maps the gene list (e.g. FlyBase gene id or Ensemble gene id) to over 40 different types of biological annotations (e.g. UniProt sequence feature, Enzyme Commission term, Gene Ontology terms, protein family information, PPI, disease associations, etc) and outputs the significantly most enriched biological annotations. We used the whole genome background for the fly and human, respectively.

4.3.3 Annotating MXE events with structural information

All the splice events were scanned against the library of CATH v4.1 FunFam HMMs (Dawson et al., 2017) using HMMER 3.0 (Eddy, 2011). DomainFinder3 (Yeats et al., 2010) was used to determine which CATH-Gene3D FunFams they belonged to. We only considered matches with a HMMER E-value of less than 0.001. For FunFams that had a known domain structure, we annotated the splice events with structure using FlyBase/Ensemble to PDB mapping. For FunFams that had no relative of known structure, we used the FunMod modelling pipeline (explained in Chapter 2) to build structural models. We used normalised DOPE and GA341 to assess the quality of the models. Only good quality models with a negative normalised DOPE score and a GA341 score of more than 0.7 were included in this analysis. For those splice events where we failed to build a model, we mapped them to the structural representative of the respective FunFams. FunFams have been demonstrated to be structurally and functionally coherent (See Chapter 2). To make sure we chose a structural representative that represent the FunFam well, the structural domain with the highest cumulative SSAP structural similarity score and the best X-ray resolution was used.

Pairs of splice isoforms were aligned to the other relatives in the FunFam using the MSA tool MAFFT (Katoh and Standley, 2013) using the function mafft-add. We extracted the alignment between the structural representative of the FunFam and the two splice isoforms. Based on this alignment, we extracted the variable residues between the two isoforms.

For those events which are not included in FunFams, we scanned them against the CATH v4.1 HMMs (to include non-FF CATH domains), SCOPe 2.06 HMMs and PDB70 June 2017 HMMs using either HMMER 3.0 or HHsearch (Remmert et al., 2012). Similarly, we only considered matches with a E-value of less than 0.001. We used FunMod modelling platform to build structural models. For those events where we failed to build a model, we mapped them to the best structural matches (based either on the HHsearch or the HMMER result). If both of the isoforms were mapped to the same structural template, we aligned the isoforms with the structural template using MAFFT and extracted the variable residues between the two isoforms.

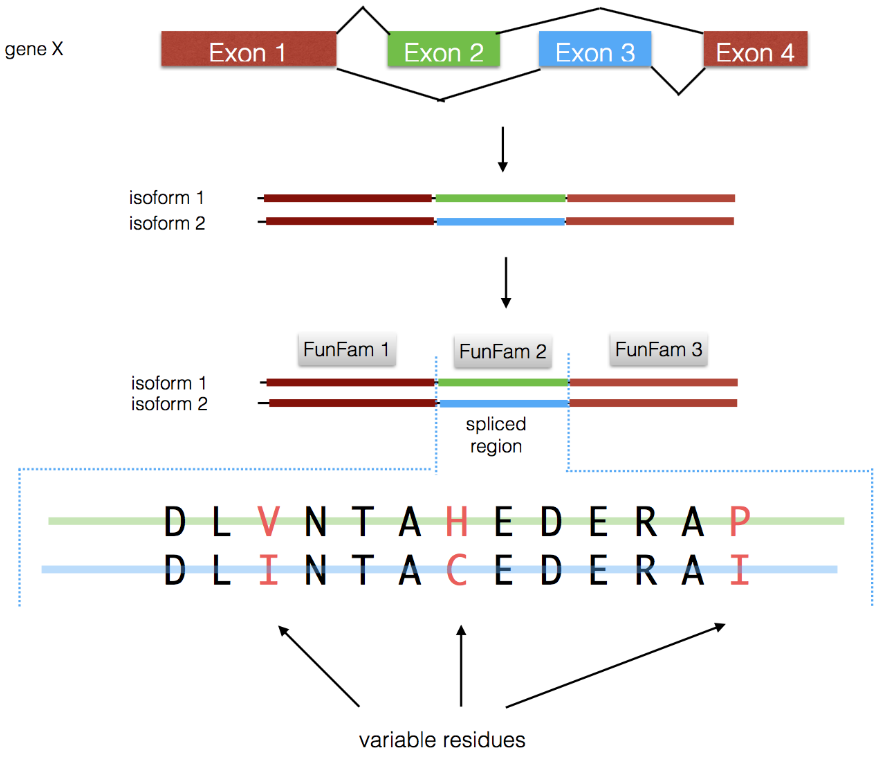


Figure 4.9: Identifying the variable residues between splice events. Note: For the cases which are not included in FunFams, the isoforms are aligned to the best structural match.

For those MXE events which we failed to annotate with structure information, we predicted if they were intrinsic disordered using IUPRED (Dosztanyi et al., 2005). IUPRED assumes that globular proteins have the capability to form many inter-residue interactions, providing the stabilising energy to overcome the entropy loss occurring during folding. In contrast, intrinsic disordered proteins do not form sufficient interactions, having a lower stabilising energy. The IUPRED predictions work by comparing the estimated pairwise interaction energy of the target protein to the pairwise interaction energy derived from a library of known structures. Residues with IUPRED score above 0.5 were considered disordered. We defined a splice isoform as intrinsically disordered if more than 50% of the residues were predicted to be disordered by IUPRED.

4.3.4 Analysis to determine if variable residues between the splice events are exposed to solvent

NACCESS (Hubbard and Thornton, 1993) is a stand-alone program that calculates the accessible area of proteins and nucleic acids in a PDB structure. This program is based on the Lee & Richards method (1971), whereby a sphere of given radius (usually 1.4 Angstroms, which is the radius of water) is rolled around the surface of the protein and the accessible surface is the total area over which the sphere passes. This is usually given as the relative accessible surface area (rASA). rASA is calculated by dividing the solvent accessible surface area with the maximum possible solvent accessible surface area for the residue.

For each splice event, we calculated the rASA for all the corresponding variable residues of the two isoforms. Amino acid residues were considered to be exposed if the rASA value was above 10% (Miller et al., 1987). For each pair of splice events we compared the solvent exposure of the complete splice region versus the solvent exposure of the variable residues. The solvent exposure was calculated as:

$$solvent\ exposure = \frac{Number\ of\ exposed\ residues}{Total\ number\ of\ residues} \tag{4.1}$$

4.3.5 Analysis to determine if variable residues between splice events, are close to functional sites

After determining whether the variable residues within the splice regions for splicing event pairs are more exposed to the solvent than the whole MXE region, we investigated if these variable residues lie in the vicinity of known functional residues. For every pair of MXE events, we determined if the variable residues within the splice regions lie close to any functional sites. We calculated the minimum distance between the atoms of the residues involved using an in-house program CalDist (developed by Dr. Andrew Martin and Camilla Pang). We used a distance cut-off of 4Å to determine if residues are close.

We also investigated random models to examine the proximity of residue to functional sites, to evaluate statistical significance. For every pair of MXE event, we created 10,000 random models (based on the patterns of location of variable residues (i.e. X-X—X-X-X, X, variable residues, -, non-variable residues), and determined the percentage of target residues that lie close to functional sites (See Figure 4.10). Then, the overall percentage of variable residues lying close to functional sites was calculated, for all MXE events and for all random events. This percentage is then compared with the actual MXE percentage. We used the z-score test to compute the statistical significance.

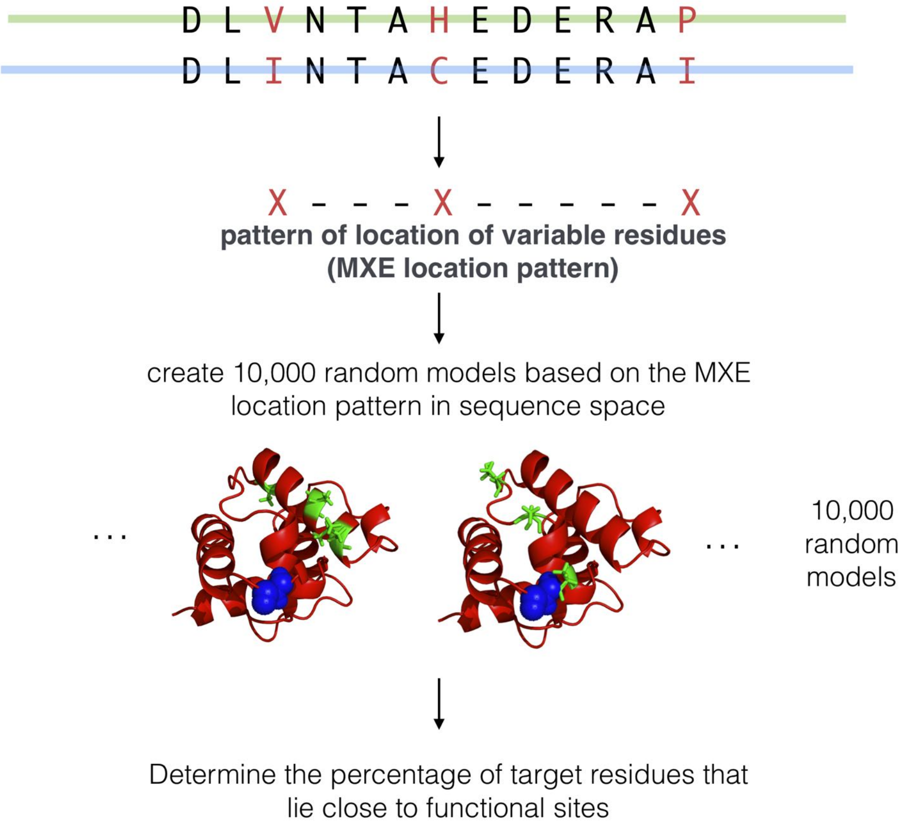

D L V N T A H E D E R A P
D L I N T A C E D E R A I
X - - - X - - - - - X
pattern of location of variable residues
(MXE location pattern)
create 10,000 random models based on the MXE
location pattern in sequence space
. . .
. . .
10,000
random
models
Determine the percentage of target residues that
lie close to functional sites

Figure 4.10: Creating random models to examine the proximity of residues to functional sites. For each pair of MXE event, we created 10,000 random models based on the pattern of location of variable residues (in sequence space). After that, we calculated the percentage of target residues that lie close to functional sites.

The functional sites considered were known experimentally characterised sites: catalytic residues taken from CSA (Furnham et al., 2014), PPI residues and protein-small molecules interaction residues (ligand binding and metal binding) taken from IBIS (Shoemaker et al., 2012). We also used our in-house predicted FunSites (Das et al., 2015; Dessailly et al., 2013) and predicted allosteric sites. FunSites are highly conserved residue positions within a Funfam. They have been found to be enriched with CSA residues, and IBIS PPI sites (Dessailly et al., 2013). For the prediction of FunSites, we only perform this analysis for FunFams that have a diverse set of relatives making it possible to distinguish between conserved positions from variable positions. The Scorecons method (Valdar, 2002) was used to calculate the sequence diversity of a FunFam by generating a diversity of position score (DOPS score). FunFams with a DOPS score of above 70 are deemed sufficiently sequence diverse for analysis (i.e. having a lower probability of predicting false positives) (Dessailly et al., 2013).

We also used the A-SITE allosteric site predictor (developed by Dr. Aurelio Moya Garcia), based on the centrality of nodes (in a protein i.e. residues) in terms of their capability to transfer signals through the protein, to predict allosteric sites in the protein, in order to determine whether variable residues lay on or close to allosteric sites.The program identifies residues with high betweenness centrality. The depletion of residues having high betweenness centrality values would be expected to interrupt the allosteric communication among regions of the protein that lie far apart. We defined the allosteric residues as the top 5 percentile residues (ranked by the A-SITE program using the betweenness centrality measure).

4.3.6 Analysis to determine if variable residues in MXE regions are clustered in 3D and lie close to functional sites

To reduce the noise and identify those variable residues likely to be having similar impacts, we clustered variable residues into structural clusters. For each pair of MXE events, we calculated the all-atom-versus-all-atom distances of the variable residues of the structural representative and determined the minimum distance between residues. We used an in-house multi-linkage clustering program which locates groups of residues that are within 8Å distance of each other. We made sure there are at least three residues present in a cluster.

Subsequently, we calculated the distance from the structural clusters to known functional residues of the structural representative. We calculated the centre-of-mass of the clusters. The centre-of-mass calculation takes the relative atomic weights into account. The minimum distance between the centre-of-mass of the cluster to the closest functional residue was calculated using an in-house program named CalDist. We used a distance cut-off of 6Å to define if the structural cluster lies close to a functional site.

We also investigated random models to examine the proximity of randomly selected residue regions to functional sites, to evaluate statistical significance. For every MXE pair, we randomly selected structural clusters that have the same size as our largest MXE structural cluster, 10,000 times, and determined the percentage of clusters that lie close to functional sites. Then, the overall percentage for all MXE events or all all random events was computed. This percentage is then compared with the actual MXE percentage. We used the z-score test to compute the statistical significance.

4.3.6.1 Analysis to determine if COSMIC cancer mutation residue clusters are close to variable residue clusters in MXE pairs

The Catalogue of Somatic Mutations in Cancer (COSMIC) is a database which collects somatic mutation from human cancer patients (Forbes et al., 2017). Somatic mutations are mutations that occur in non-germ line cells i.e. which are not passed on to human offspring. Accumulation of these mutations may alter cellular functions and contribute to cancer (Martincorena and Campbell, 2015). COSMIC mutation data are derived either from the literature, whole genome sequencing data or extracted from databases such as the Cancer Genome Atlas and the International Cancer Genome Consortium.

A number of studies have that found that cancer mutation sites lie in the vicinity of functional residues (i.e. protein and ligand binding sites) (David and Sternberg, 2015; Jubb et al., 2016; Yamada et al., 2016), therefore we determined if COSMIC cancer mutation residues/clusters lie close to MXE variable residues and functional residues. We mapped the COSMIC cancer mutation residues (version 80, released on Feb 2017) onto the human MXE events. For those where we have structural information, we determined if the variable residue residues/clusters are close to COSMIC residues/functional sites using a distance cut-off of 4Å (per-residue) or 6Å (per-cluster).

4.3.7 Quantifying the residue changes using a physiochemical score

We also compared the physiochemical properties of variable residues using the McLachlan physiochemical similarity matrix (McLachlan, 1972). This scoring matrix provides a measure of amino acid similarity based upon amino acid polarity, size, shape and charge. Depending on the similarity in these characteristics, a pair of amino acids was given a similarity score ranging from zero to six. A score of zero indicates no similarity or a deletion. The score for a pair of identical amino acids is typically five, but six for amino acids considered less common by McLachlan (1972). These less common amino acids consisted of phenylalanine, methionine, tyrosine, histidine, cysteine, tryptophan, arginine, and glycine (McLachlan, 1972).

Because we are interested in whether the variable residues exhibit changes in physiochemical properties, we transformed the McLachlan scores: 1 being very similar and 7 being very dissimilar. We summed up all the Mclachlan similarity scores for a set of variable residues and divided by the total number of variable residues.

4.4 Results and Discussion

4.4.1 Statistics of the MXE dataset

The fly MXE dataset consisted of 416 splice isoform pairs that involved a total of 118 fly genes. There were a total of 1,917 fly MXE pairs identified. Out of the 1,917 MXE pairs, 1,643 pairs involved the DSCAM gene and 273 involved non-DSCAM genes. The fly DSCAM genes have a total of 4 variable exon clusters giving 38,016 potential splicing isoforms. This gene is important for fly immunity and neuronal processes. We decided to split the MXE events into DSCAM events and non-DSCAM events to remove bias. In contrast, the human MXE dataset consisted of 460 splice isoforms that involved a total of 205 genes. There were a total of 241 human MXE pairs identified. Figure 4.11 demonstrates the length of the MXEs. They span from 10 residues up to 400 residues, with an average size of around 70.

We used DAVID, a function annotation tool, to identify the enriched biological annotations for these MXE genes realativerelative to the whole genome background. The MXE genes are associated with membrane proteins (e.g. ion channels, synaptic vesicles and receptors) and enzymes (e.g. kinases, transferases, hydrolases and oxidoreductases) with false discovery rate level <0.01.

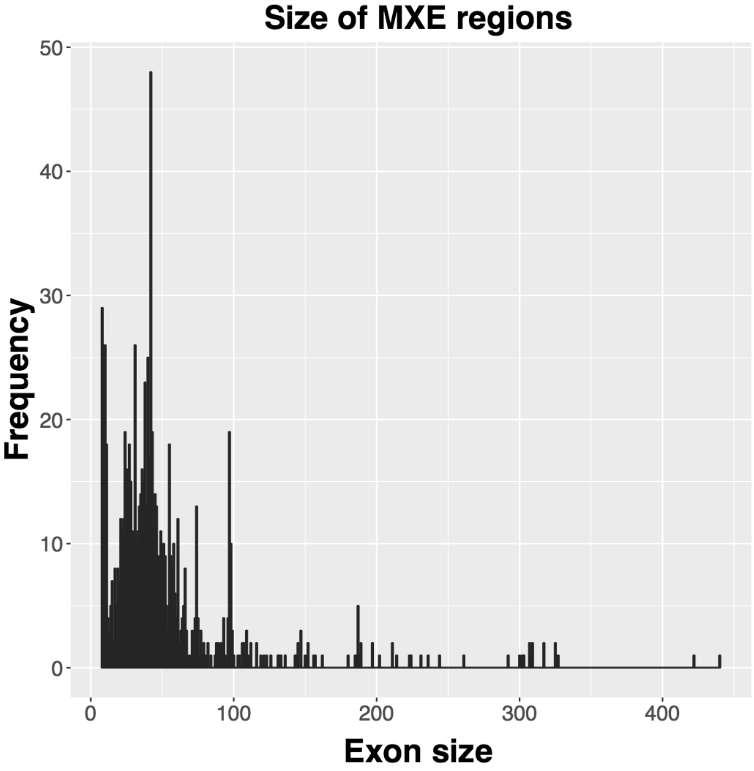


Figure 4.11: Size of the MXE regions. The exon size is represented by the number of amino acids.

We structurally annotated the splice isoforms with known protein structures from the PDB. Structural models were built using the in-house FunMod modelling platform (See Chapter 2). For those events where we failed to produce a good model, we performed structural mapping to the representative structure of the respective FunFam. We annotated the whole protein sequences where possible, however, if we failed to do so, we made sure the splice region was covered. All 1,643 pairs of splicing events for DSCAM genes could be mapped to FunFams. 145 out of the 274 pairs of non-DSCAM splicing events could be mapped to a FunFam. Only 63 pairs out of the 241 pairs of human MXE events could be mapped to a FunFam.

For those events that we failed to mapped to a FunFam, we performed structural modelling and structural mapping based on non-FunFam CATH domains, SCOPe domains or PDB chains. We also performed disorder analysis using the program IUPRED. We defined a splice isoform as disordered if more than 50% of the residues inside the slice were predicted to be disordered by IUPRED. Figure 4.12 gives information on structural annotations for the sets of splice isoforms. We managed to annotate 75% of fly isoforms and 50% of the human isoforms with structural information. Less than 10% of the isoforms can be mapped to known structures. 5% of the fly isoforms and 18% of the human isoforms are disordered.

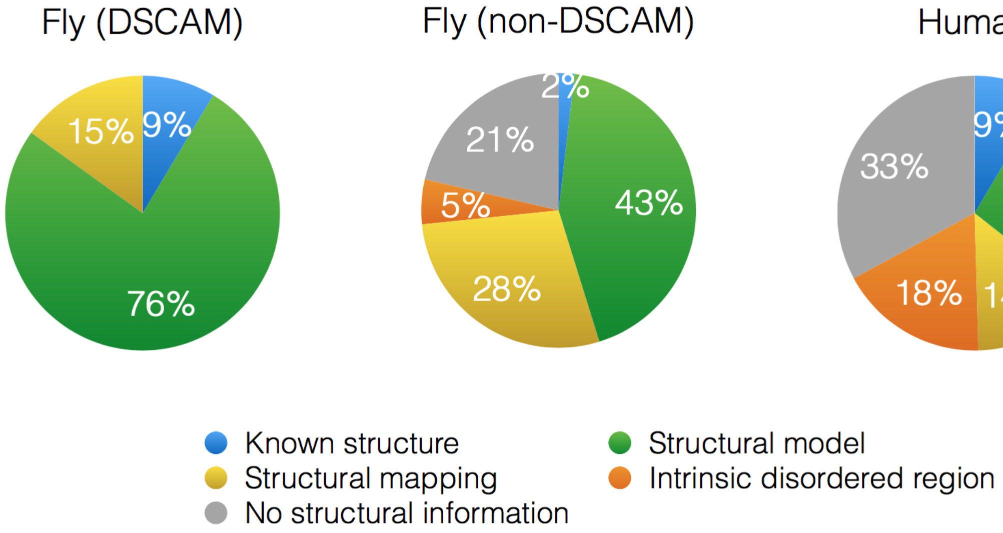


Figure 4.12: Splice isoforms with structural information.

We also calculated the proportion of the FunFam domain covered by the splice region. The distribution of coverage was quite large with a mean value of 36% (See Figure 4.13). We also determined the number of variable residues involved in MXE events. They usually involved less than 10 residues (See Figure 4.14).

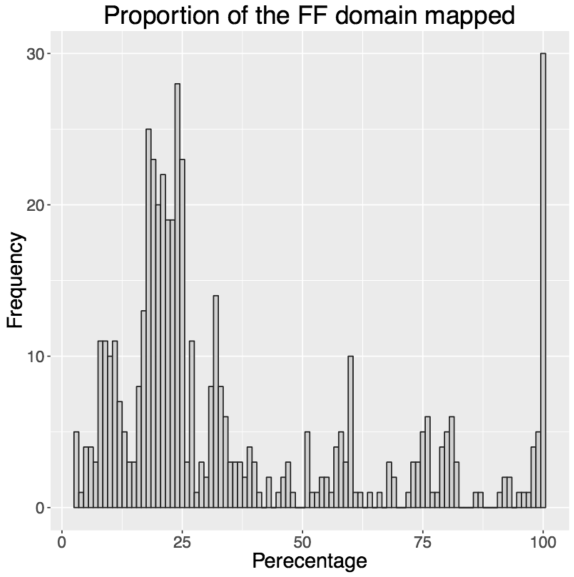


Figure 4.13: Proportion of the FunFam domain mapped by the splice region

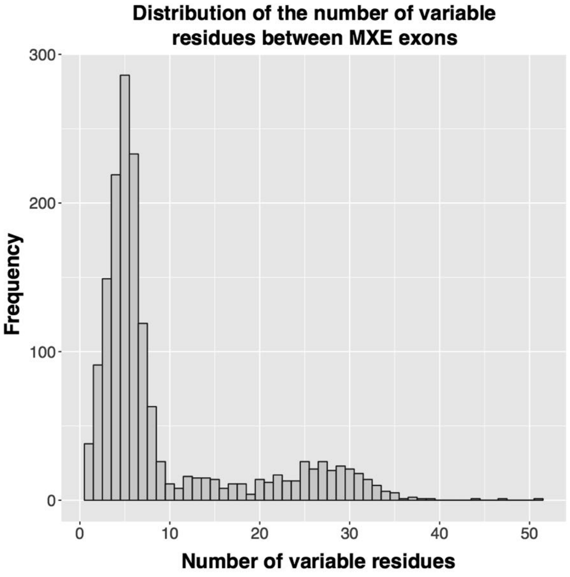


Figure 4.14: Number of variable residues between MXE exons

Figure 4.15 illustrates the distribution of FunFams containing the MXE splice region for DSCAM. All the splice regions for DSCAM events are mapped to the CATH superfamily 2.60.40.10 (the Immunoglobulins-like superfamily). This superfamily is one of the most highly diverse CATH superfamilies with 19,468 domains. The splicing events were mapped to six FunFams within the CATH superfamily. For fly non-DSCAM and human MXE events, the FunFams mapped by the splice region are distributed among 84 CATH superfamilies (See Figure 4.16 and 4.17). Figure 4.18 shows the number of CATH superfamilies and FunFams common to both fly and human. Figure 4.19 lists the names of common FunFams. These are generally membrane proteins associated with cell-cell adhesion, signal transduction and molecule/ion transport.

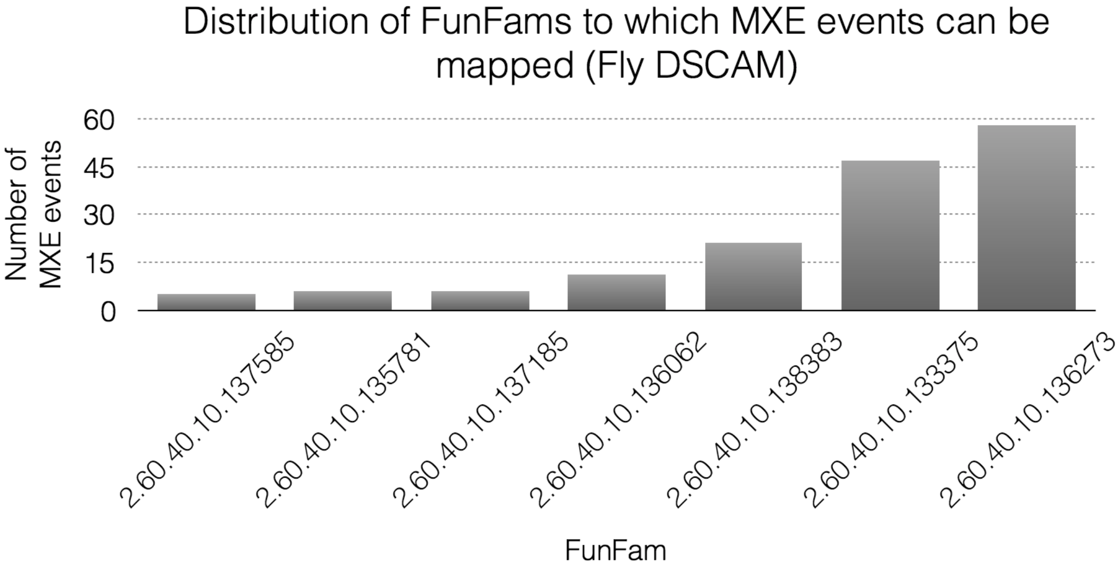


Figure 4.15: Distribution of FunFams to which MXE events can be mapped (Fly, DSCAM)

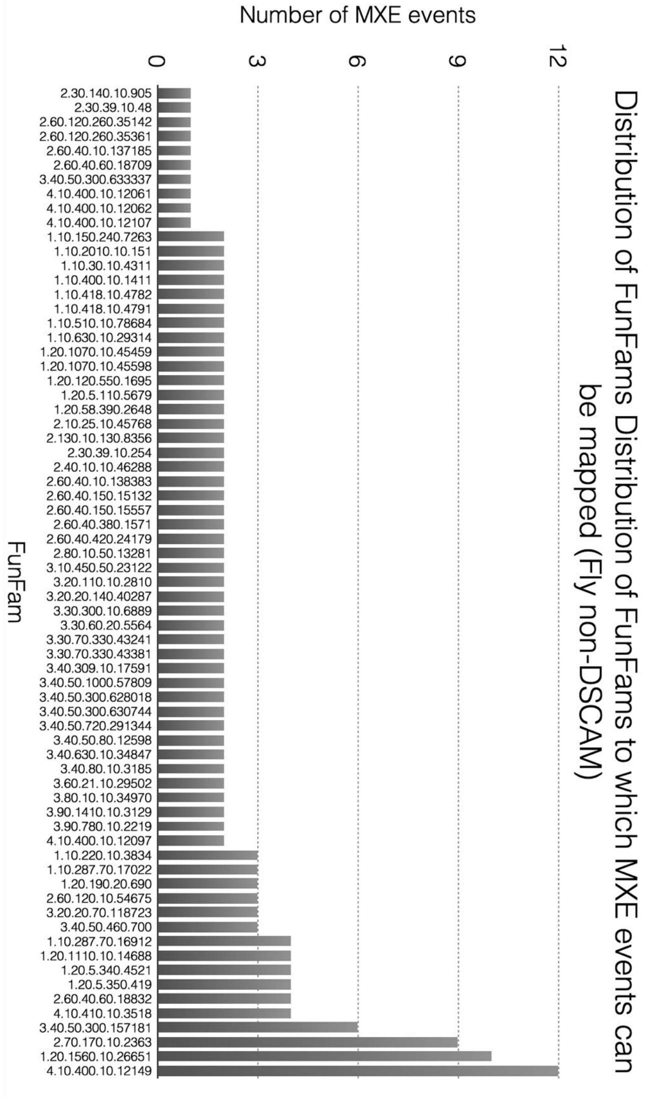


Figure 4.16: Distribution of FunFams to which MXE events can be mapped (Fly, nonDSCAM)

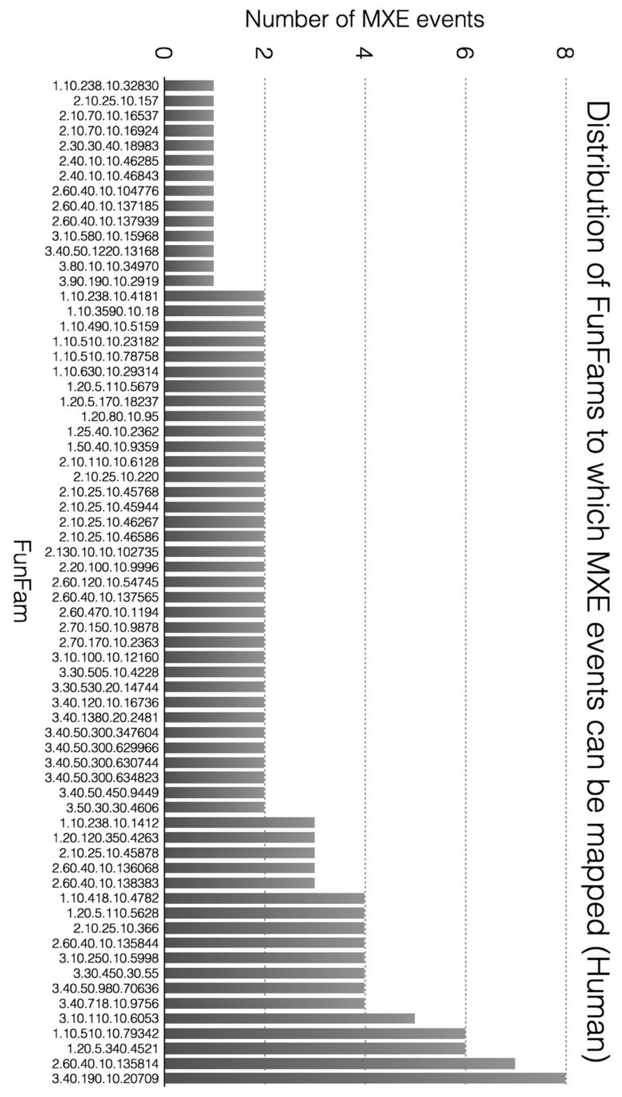


Figure 4.17: Distribution of FunFams to which MXE events can be mapped (Human)

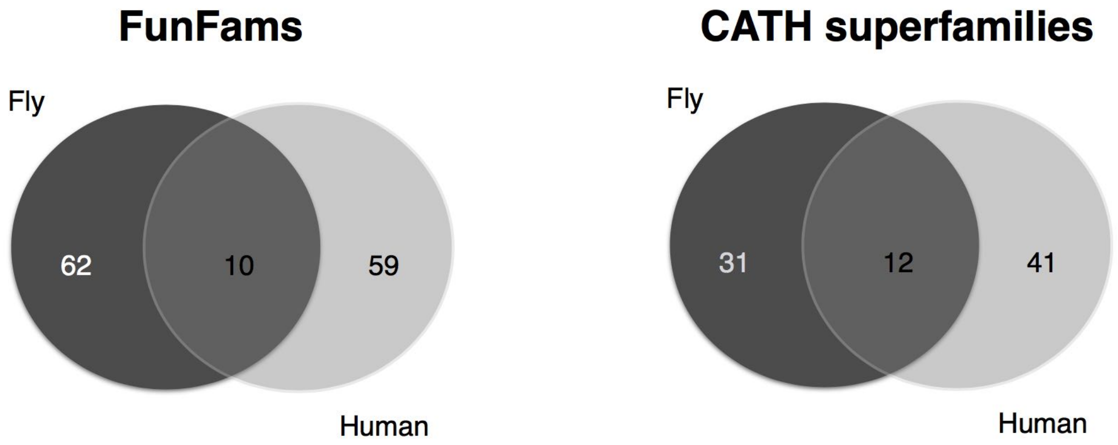
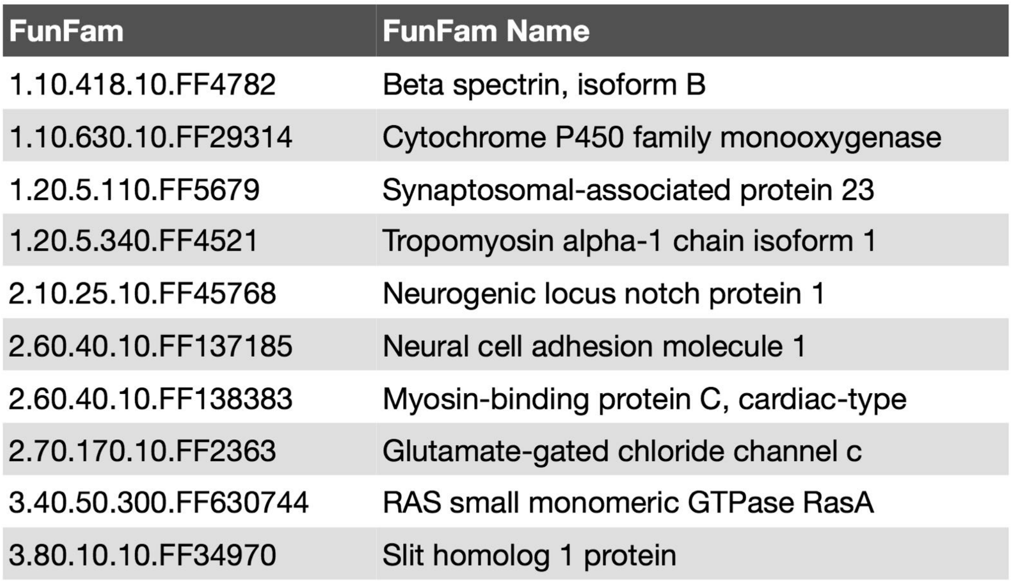


Figure 4.18: Number of CATH superfamilies and FunFams common to both fly and human.



FunFam	FunFam Name
1.10.418.10.FF4782	Beta spectrin, isoform B
1.10.630.10.FF29314	Cytochrome P450 family monooxygenase
1.20.5.110.FF5679	Synaptosomal-associated protein 23
1.20.5.340.FF4521	Tropomyosin alpha-1 chain isoform 1
2.10.25.10.FF45768	Neurogenic locus notch protein 1
2.60.40.10.FF137185	Neural cell adhesion molecule 1
2.60.40.10.FF138383	Myosin-binding protein C, cardiac-type
2.70.170.10.FF2363	Glutamate-gated chloride channel c
3.40.50.300.FF630744	RAS small monomeric GTPase RasA
3.80.10.10.FF34970	Slit homolog 1 protein

Figure 4.19: Names of common FunFams to both fly and human. They are generally membrane proteins associated with cell-cell adhesion, signal transduction and molecule/ion transport.

4.4.2 Analysing the frequency with which MXE events are exposed to solvent

We determined whether the variable residues involved in MXE events are more exposed to the solvent than the whole spliced region. We calculated the rASAs of variable residues using the NACCESS program for all pairs of splicing events. For comparison, we also calculated the rASAs of all mapped residues (all the residues found within the splice region). We considered all residues that have an rASA value of >10% to be exposed. The percentage of exposed residues was calculated by dividing the number of exposed residues by the total number of mapped residues. Figures 4.20, 4.21, 4.22 compares the proportion of exposed variable residues to the proportion of all exposed MXE residues for fly DSCAM, fly non-DSCAM, and Human events. Overall, we can observe that variable residues are more exposed than all MXE residues. The difference was highly significant (p-value <2.2e-16, Wilcoxon signed rank test).

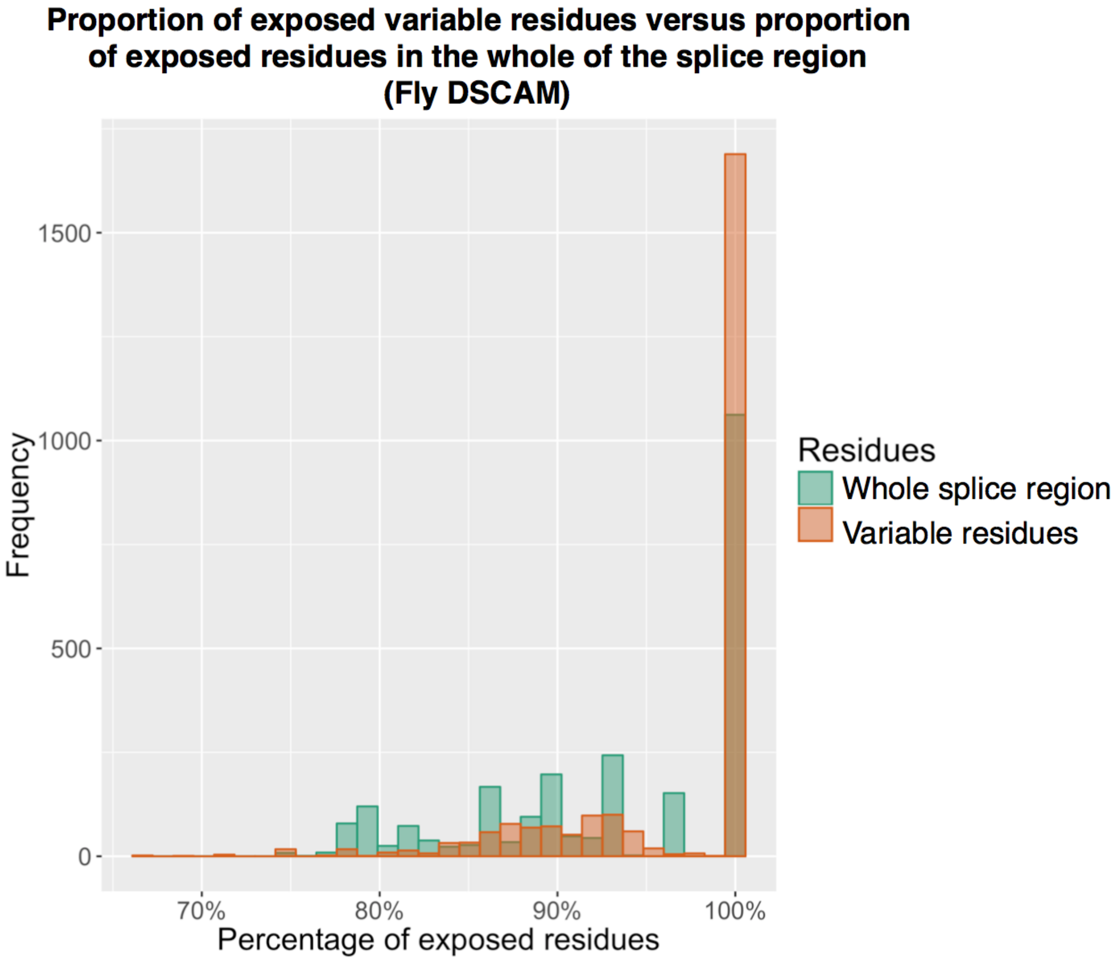


Figure 4.20: Proportion of exposed variable residues versus exposed residues in the whole of the splice region (Fly, DSCAM)

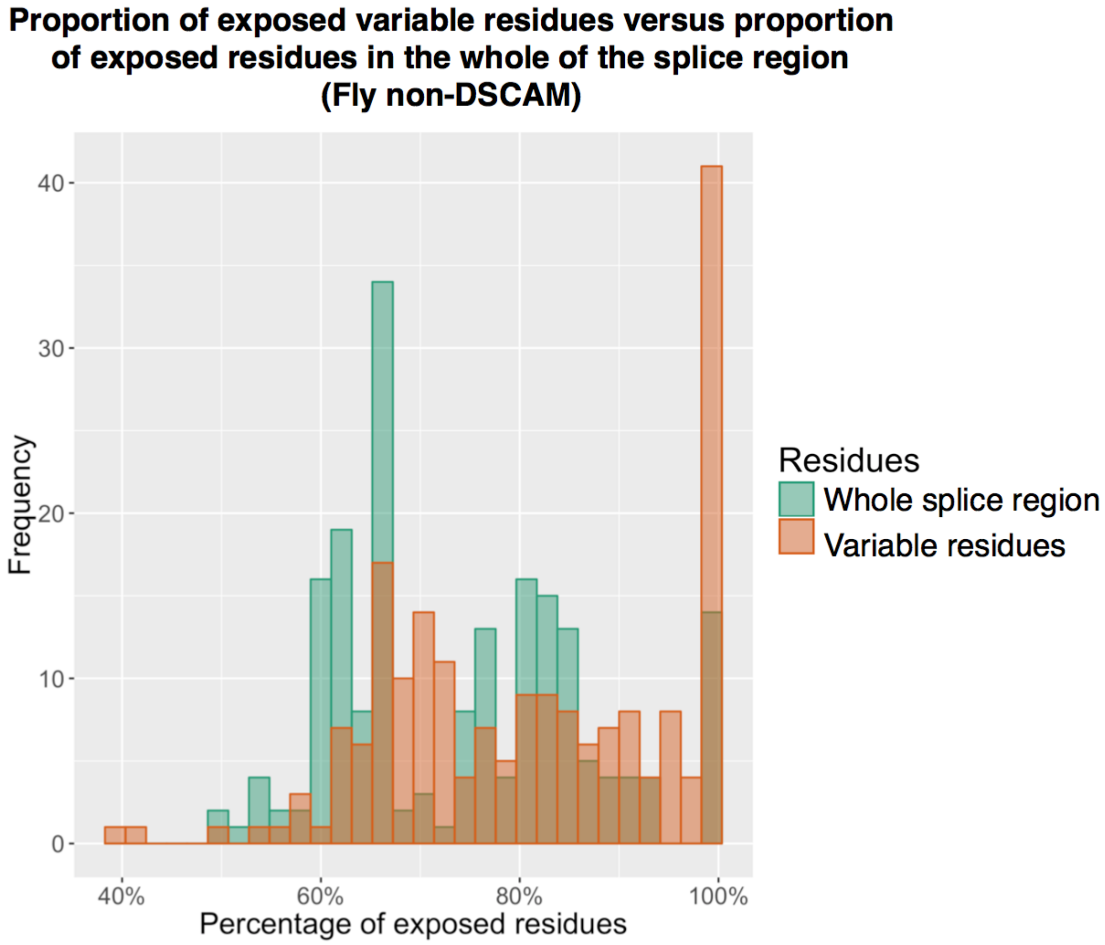


Figure 4.21: Proportion of exposed variable residues versus exposed residues in the whole of the splice region (Fly, non-DSCAM)

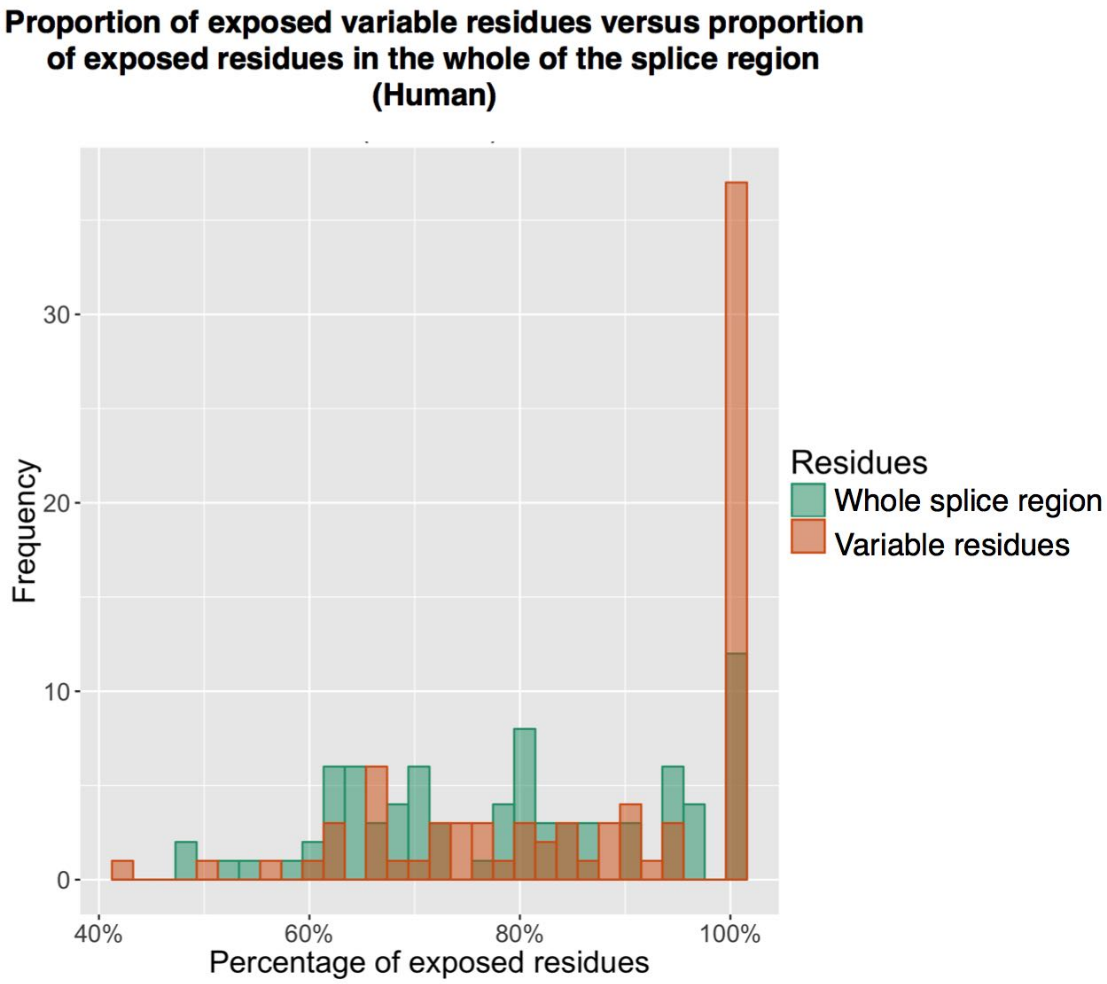


Figure 4.22: Proportion of exposed variable residues versus exposed residues in the whole of the splice region (Human)

4.4.3 How often do variable residues in MXE events lie close to functional sites

In the previous section, we demonstrated that the variable residues within the splice regions for splicing event pairs are more exposed to the solvent than the whole MXE region. Here, we investigate if these variable residues lie in the vicinity of known functional residues. The known functional residues considered were catalytic residues (obtained from CSA), PPI sites (obtained from IBIS database), protein-small molecules interaction sites (ligand binding sites, obtained from IBIS database). We also considered our in-house functional sites (FunSites, enriched with catalytic sites and PPI sites) and allosteric sites (based on the between centrality measure).

For every pair of splice isoforms compared, we determined if the variable residues within the splice regions are close to any functional sites. We calculated the minimum distance between the atoms of the residues involved using an in-house program CalDist. We used a distance cut-off of 4Å to determine if residues are close. The results were compared with a random model - described in Methods.

Figure 4.23 shows that there is a statistically significant tendency for MXE events with variable residues to lie close to PPI, protein-small molecule and FunSites for both fly and human dataset. For Fly non-DSCAM events, there is also a tendency for variable residues to be in the vicinity of allosteric sites. We had less 10 events annotated with catalytic sites but these did not usually lie close to a MXE region using 4Å distance cutoff.

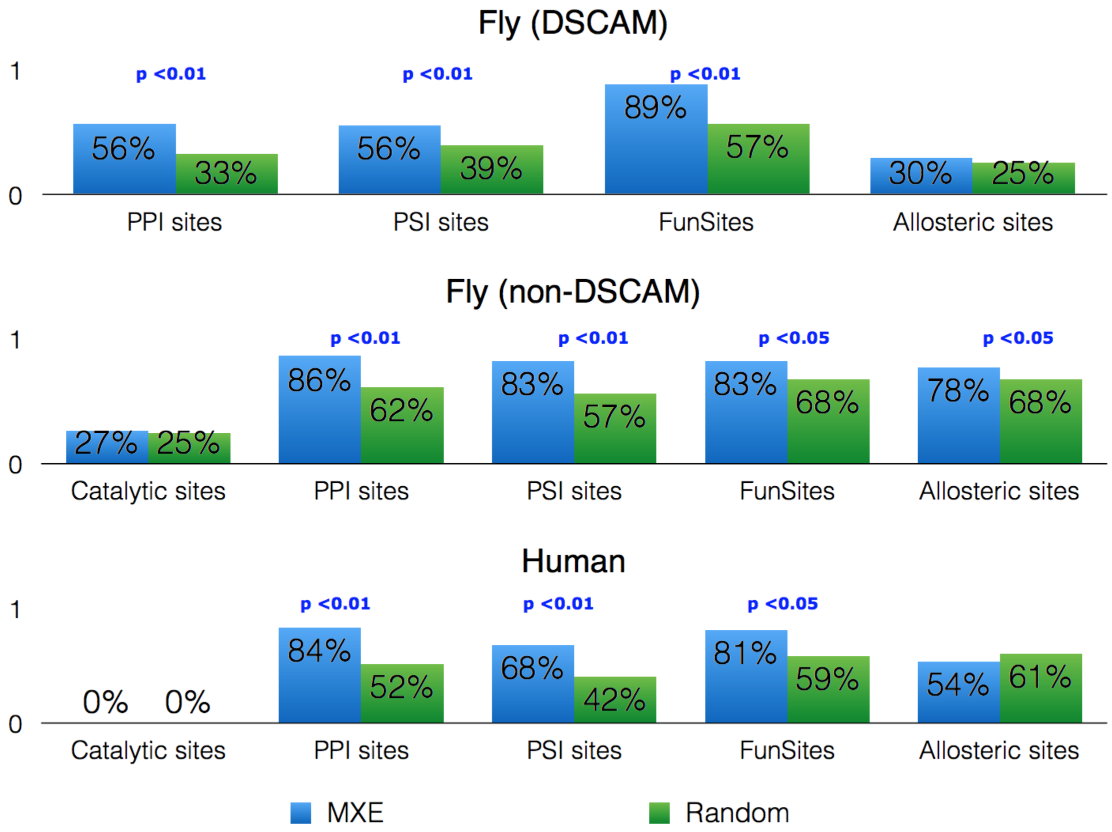


Figure 4.23: Proportion of variable residues in MXE events that lie close to functional sites compared to the number of random residues that lie close to functional sites.

4.4.4 How often do clusters of variable residues in MXE events lie close to functional sites

In order to remove noise, we structurally clustered the variable residues. Clusters of variable residues were formed in 91% of the fly non-DSCAM events, 88% of the fly DSCAM events and 73% of the human MXE events. For those events that formed clusters, about 61% of the fly non-DSCAM events, about 50% of the fly DSCAM events and about 66% of the human events formed only one cluster. For those that formed multiple clusters, they usually formed 2-3 clusters (See Figure 4.24). We also determined the size of the individual clusters for events with multiple clusters. There's usually a single large cluster followed by smaller clusters (See Figure 4.25).

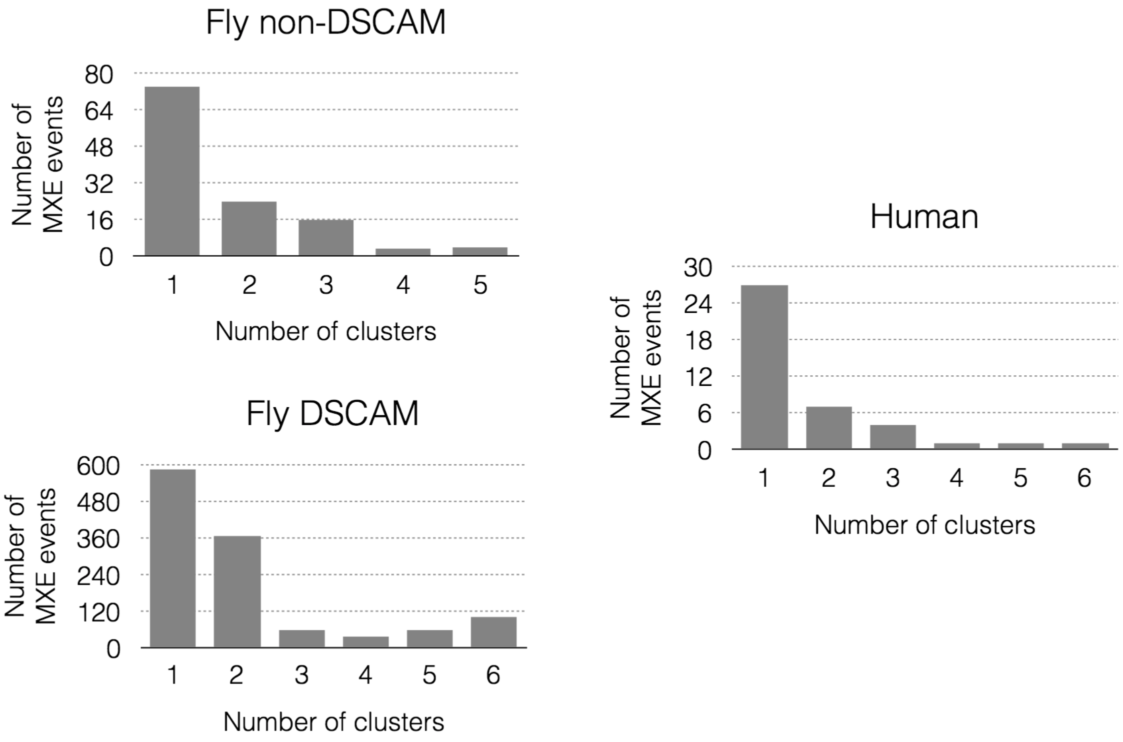


Figure 4.24: Number of structural clusters formed by MXE variable residues.

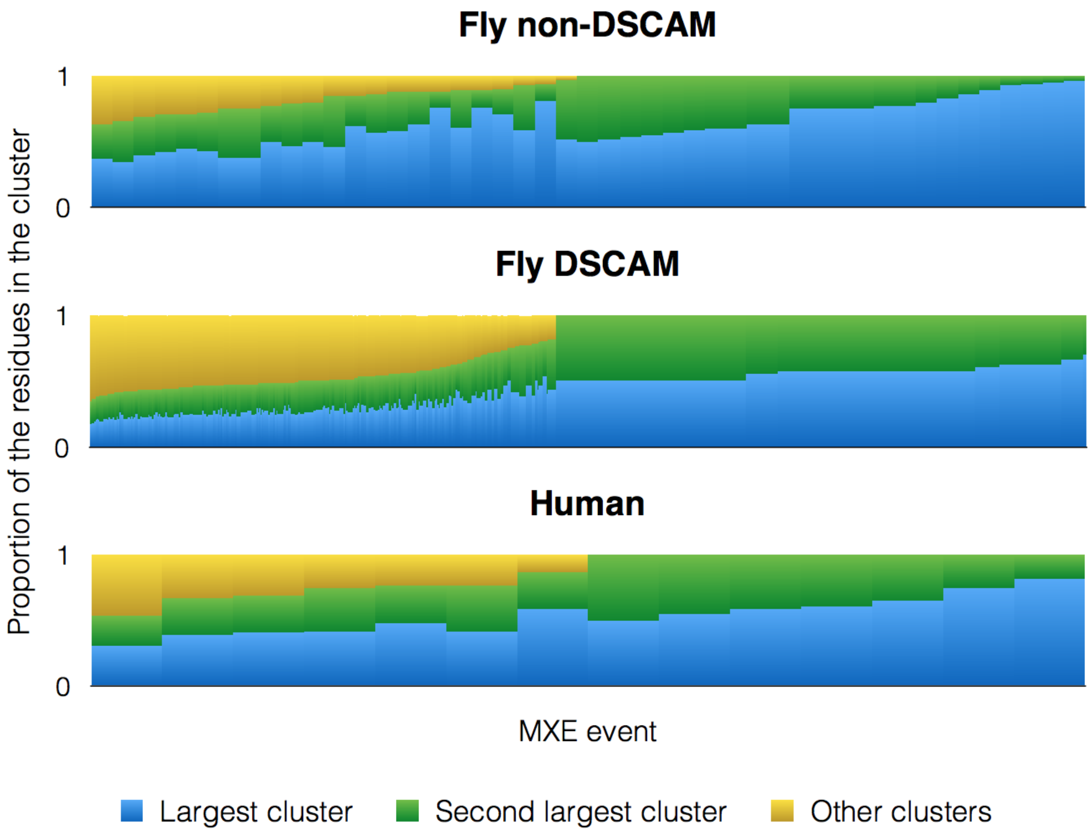


Figure 4.25: Size of structural clusters formed (only showing events with multiple clusters) for variable residues in MXE events

Next, we determined whether structural clusters lie close to functional residues. For every pair of splice isoforms compared, we determined if centre of mass of the MXE cluster was close to any functional sites (within 6Å). We calculated the minimum distance between the atoms of the residues involved using an in-house program CalDist. The results were compared with a random model - described in Methods.

Figure 4.26 shows that there is a statistically significant tendency for variable region clusters of MXE events to lie close to PPI, protein-small molecule, FunSites and allosteric sites for both fly and human datasets. When examining the likelihood for variable residue clusters to lie close to FunSites, for human MXE events, the p-value is not significant due to the low sample size (i.e. our functional site prediction protocol was not able to predict many sites because the FunFam involved have low DOPS score (<70, See Methods)).

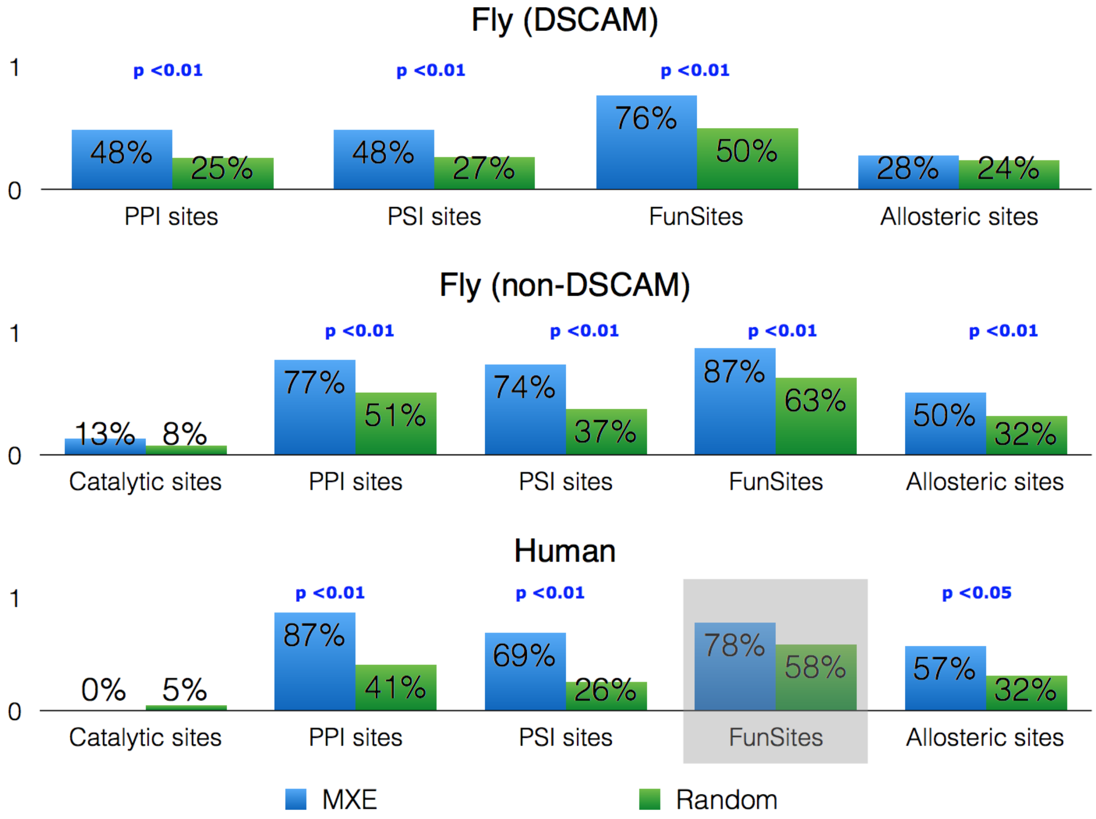


Figure 4.26: Proportion of variable region clusters in MXE events that lie close to functional sites compared to the number of random clusters that lie close to functional sites. NOTE: The number of human MXE events with FunSite predictions is small.

4.4.4.1 How often do variable residues/clusters in MXE events lie close to putative cancer mutation sites

A number of studies that have functional residues (i.e. protein and ligand binding sites) to be in the vicinity of cancer mutation sites (David and Sternberg, 2015; Jubb et al., 2016; Yamada et al., 2016). We therefore we investigated whether if COSMIC cancer mutation residues coincide with both MXE variable residues and functional residues. Putative cancer mutations from the COSMIC database were mapped to 44 of our human MXE events. We annotated 30 of these events with structures. We then determined if the variable residues/clusters in these MXE events were close to COSMIC mutation residues. We found that 23 of the MXE events were close to COSMIC residues (statistically significant with e-value <0.01, compared with random models). We obtained the same results for both per-residue and per-cluster analyses.

We checked if these events were also close to functional residues and found 21 of the events were close.A significant number of cosmic mutations seems to be affecting the MXE clusters that are close to functional sites, suggests cancer mutation could be affecting function associated with the MXE functional switching events. Section 4.4.6.3 below, presents an example where the variable residues coincide with both COSMIC mutation residue and known PPI sites.

4.4.5 How often do MXE events involve a significant change in physiochemical properties of the variable residues

4.4.5.1 Variable residues close to functional sites

For MXE events where variable residues lie close (<4Å) to functional sites, we compared the physiochemical properties of the equivalent residues between the two isoforms using McLachlan's physiochemical matrix. We transformed the McLachlan scores such that the higher the McLachlan score, the more dissimilar the two residues. We summed up all the McLachlan similarity scores (i.e. for every pair of variable residues) and divided by the total number of variable residues examined, to give an average score for the splice pair we were comparing.

For the fly non-DSCAM dataset, there are significant physiochemical changes in residue properties for more than 78% of the variable regions close to FunSites, PPI sites, PSI sites and allosteric sites, which have average McLachlan scores above 5.5 (See Figure 4.27).

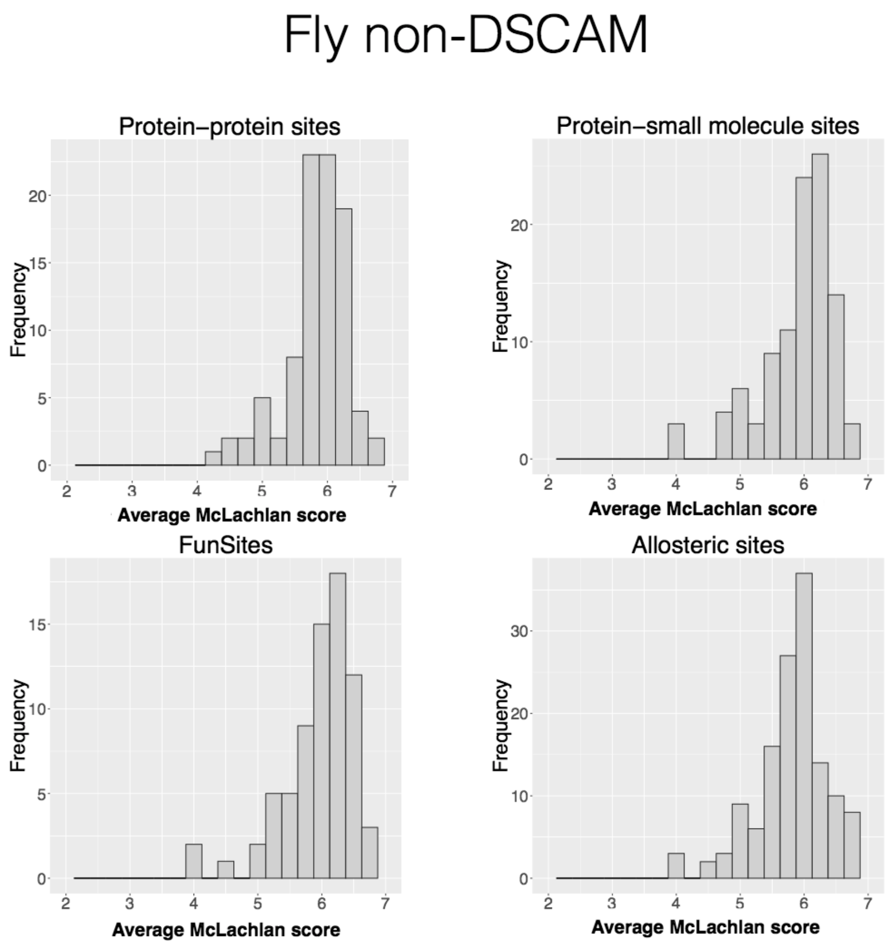


Figure 4.27: Distribution of average McLachlan scores for changes in variable MXE residues located close (<4Å) to the different types of functional sites (Fly,non-DSCAM)

For the fly DSCAM dataset, more than 79% of variable region changes close to allosteric sites have average Mclachlan scores above 5.5 (See Figure 4.28). Whilst 65% of variable region changes close to ligand binding and interfaces, and 57% of variable region changes close to FunSites have high average Mclachlan scores (>5.5). Thus, fly DSCAM MXE events are more likely to involve a profound change for residues lying close to allosteric sites than interfaces and ligand binding sites.

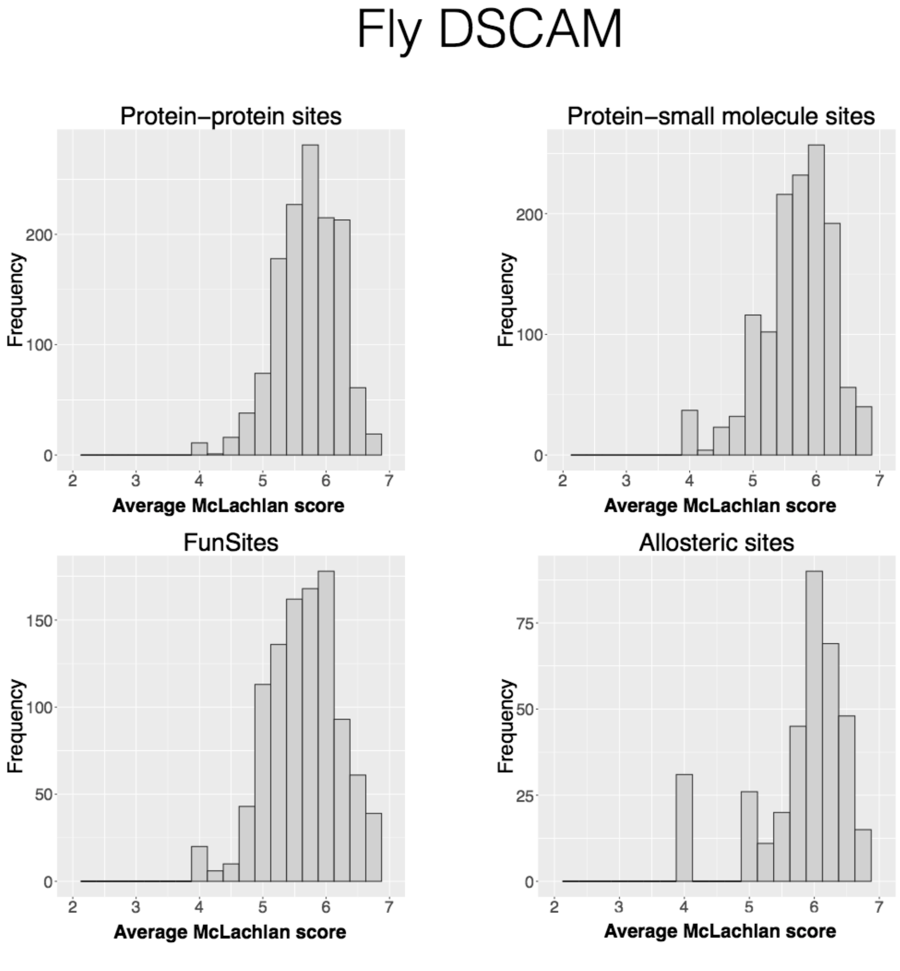


Figure 4.28: Distribution of average Mclachlan scores for changes in variable MXE residues located close (<4Å) to the different types of functional sites (Fly, DSCAM)

Similarly to the fly non-DSCAM dataset, more than 74% of the variable region changes for FunSites, PPI sites, PSI sites and Allosteric sites have average McLachlan scores of above 5.5 for the human dataset(See Figure 4.29). Thus, MXE events which can be mapped to structure show a significant tendency to lie close to a known and/or predicted functional site and if they do they are likely to be associated with a significant change in their physiochemical property.

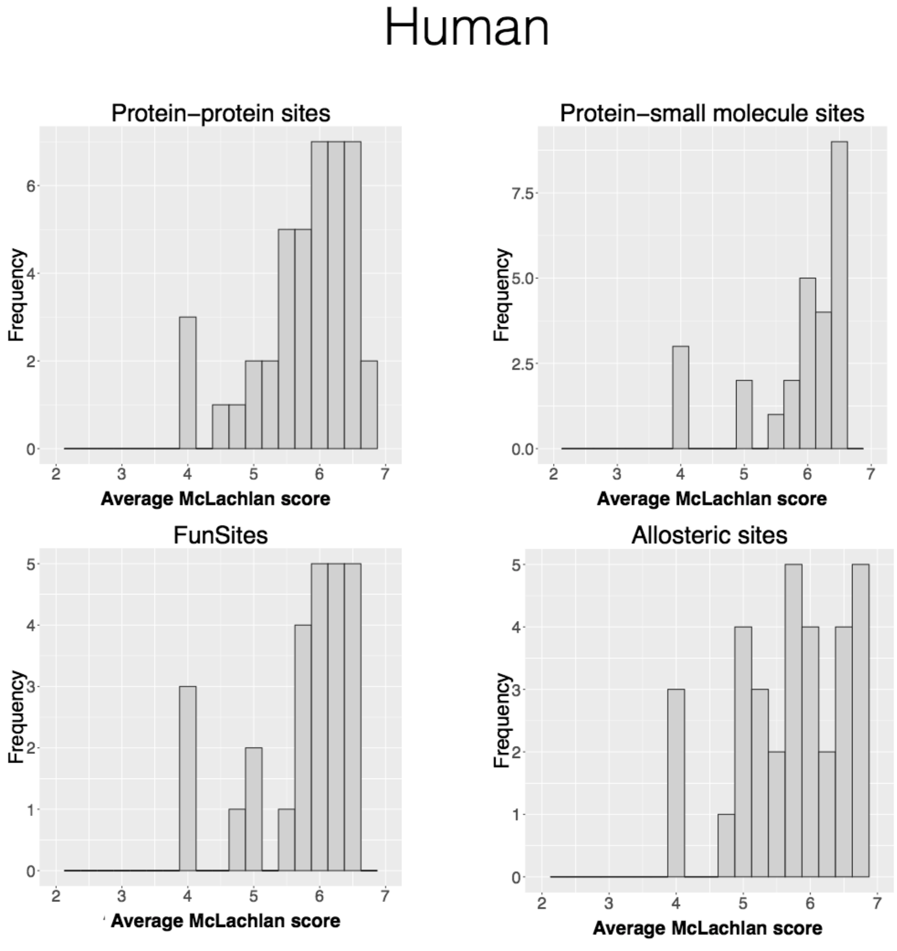


Figure 4.29: Distribution of average McLachlan scores for changes in variable MXE residues located close (<4Å) to the different types of functional sites (Human)

4.4.5.2 Physiochemical changes of Variable residues not located close to functional sites

We also calculated the physiochemical property of variable residues not located close to functional sites. We obtained result similar results to those for variable residues that are close to functional sites, with most of them undergoing significant physiochemical changes (average McLachlan scores >5.5, See Figure 4.30).

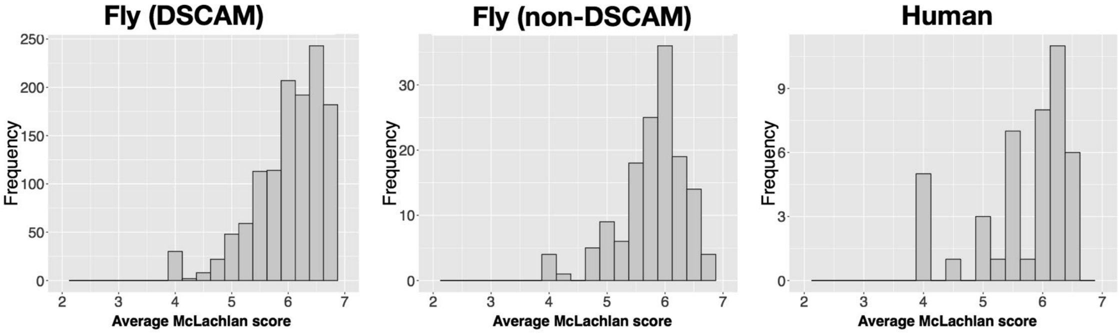


Figure 4.30: Distribution of average McLachlan scores for changes in variable MXE residues more distant from functional sites

We examined whether variable residues that are closer to functional sites ($<$4Å) have more significant changes in physiochemical properties than those that are distant. Using a strigent cutoff of 6, for the fly non-DSCAM and human dataset, there is a higher proportion of MXE events with McLachlan score $>$6 for variable residues that are closer to functional sites, but this is not statistically significant. We obtained a conflicting result for the Fly DSCAM dataset where variable residues that are distant to functional sites have a higher proportion of MXE events with McLachlan score $>$6.

Overall, these results indicate that in general the MXE events produce significant changes in physiochemical properties of the MXE regions, regardless of whether they are close to functional sites or not.

4.4.6 Some examples of proteins displaying likely functional shifts following MXE event

In this section, we present a few examples to illustrate the impacts of MXE events on the function of the proteins.

4.4.6.1 Collapsin Response Mediator Protein (CRMP-PA versus CRMP-PB)

The Collapsin Response Mediator Protein (FBgn0023023) is a Dihydropyrimidinase enzyme (EC 3.5.2.2) that catalyzes the chemical reaction:

$$5,6-dihydrouracil+H(2)O <=> 3-ureidopropanoate. \tag{4.2}$$

It is predicted to be involved in biological processes such as centrosome localization, pyrimidine nucleobase biosynthetic process, and positive regulation of Notch signaling pathway.

Currently, there is no experimental structure for the fly isoforms (CRMP-PA and CRMP-PB). The MXE exons can be mapped to FunFam (3.20.20.140.FF40287). This FunFam is populated with Dihydropyrimidinase sequences. We modelled both of the isoforms using the same template (CATH domain 1gkpC02) from the FunFam. The models generated have a good model with GA341 score of 1 and normalised DOPE score of -0.71 suggesting they are of good quality. The structures share a high structural similarity (SSAP Score above 91) and sequence similarity (79%) (See Figure 4.31 for a structural superposition with the splice regions highlighted).

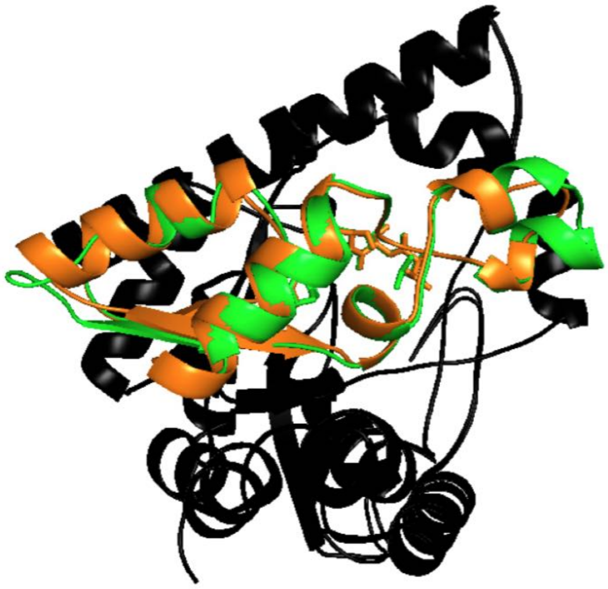


Figure 4.31: The splice regions of the protein isoforms CRMP-PA(orange) and CRMP-PB (green)

We used the program APBS to calculate surface potentials of the two isoforms. Adaptive Poisson-Boltzmann Solver (APBS) is a software package which models biomolecular solvation by solving the Poisson-Boltzmann equation (PBE) (Baker et al., 2001). PBE is a popular continuum model used to describe electrostatic interactions between solutes in salty, aqueous media. We can see from Figure 4.32 that the surface potential differs between the two isoforms. Over 80% of exposed splice region in CRMP-PA is positively charged. In contrast, the surface potential for CRMP-PB has a more negative charge that covers 60% of the exposed splice region.

We found out that there is a change in the zinc-binding residue 192, with a histidine in isoform CRMP-PA and a glycine is found in isoform CRMP-PB. This change is likely to disrupt the binding of the zinc molecule in the CRMP-PB isoform (See Figure 4.33).

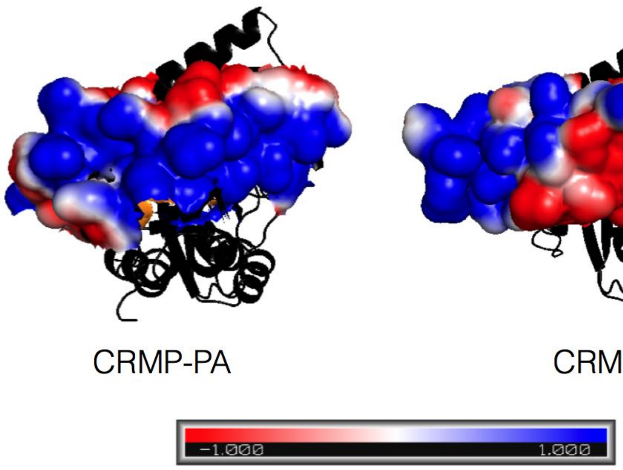


Figure 4.32: The surface potential of the splice regions (CRMP-PA and CRMP-PB)

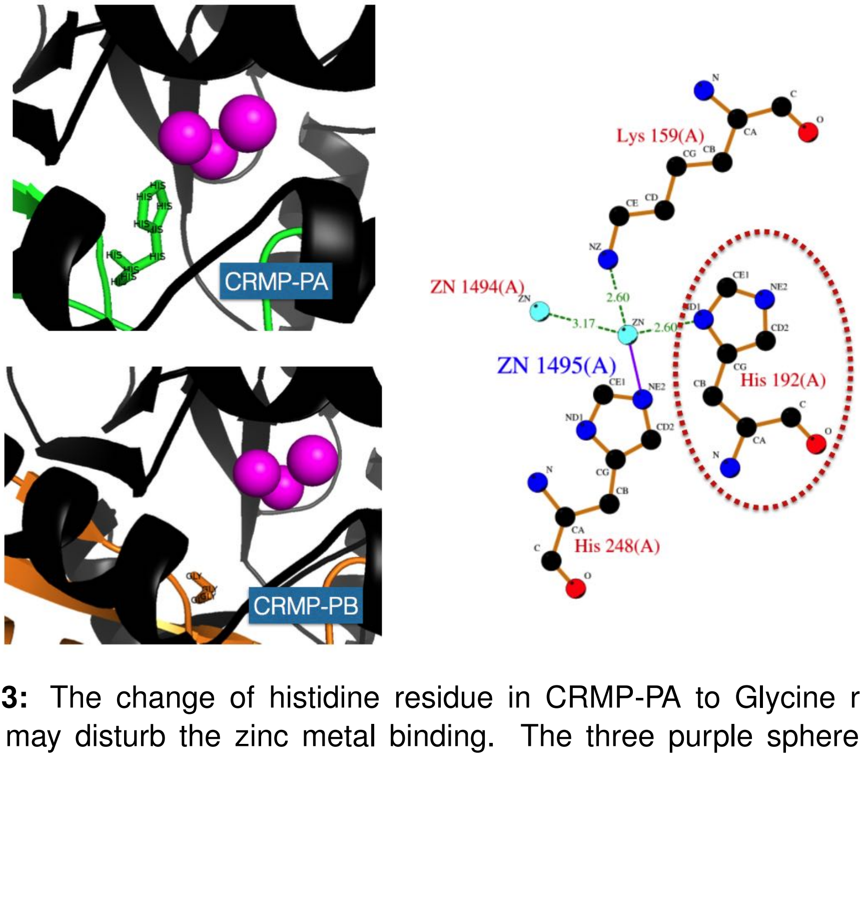
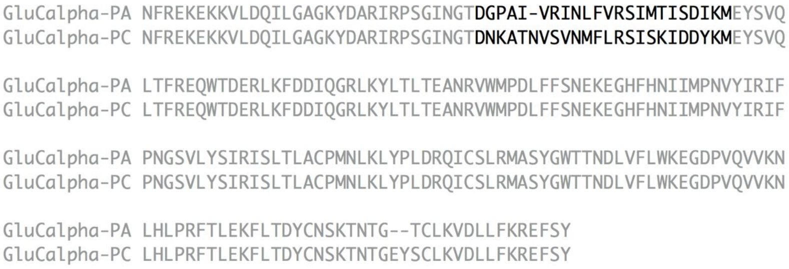


Figure 4.33: The change of histidine residue in CRMP-PA to Glycine residue in CRMP-PB may disturb the zinc metal binding. The three purple spheres are the zinc ions.

4.4.6.2 Glutamate C-alpha protein (Glucaplha-PA versus Glucaplha-PC)

The Glutamate C-alpha protein is a glutamate-gated chloride channel. As the protein name suggests, it is a ion channel gate that mediates chloride transport. As described in UniProt, the protein has a main extracellular domain of about 200 residues, followed by 4 transmembrane helices. The MXE events occur in the extracellular region. They are mapped to the Glutamate-gated chloride channel c FunFam (2.70.170.10.FF2363). We extracted the segments of the two isoforms, which are mapped to the FunFam, and performed a sequence alignment between them. (See Figure 4.34) There are about 24 residues that make up the splice region.



```
GluCalpha-PA NFREKEKKVLDQILGAGKYDARIRPSGINGTDGPAI-VRINLFVRSIMTISDIKMEYSVQ
GluCalpha-PC NFREKEKKVLDQILGAGKYDARIRPSGINGTDNKATNVSVNMFLRSISKIDDYKMEYSVQ

GluCalpha-PA LTFREQWTDERLKFDDIQGRLKYLTLTEANRVWMPDLFFSNEKEGHFHNIIMPNVYIRIF
GluCalpha-PC LTFREQWTDERLKFDDIQGRLKYLTLTEANRVWMPDLFFSNEKEGHFHNIIMPNVYIRIF

GluCalpha-PA PNGSVLYSIRISLTLACPMNLKLYPLDRQICSLRMASYGWTTNDLVFLWKEGDPVQVVKN
GluCalpha-PC PNGSVLYSIRISLTLACPMNLKLYPLDRQICSLRMASYGWTTNDLVFLWKEGDPVQVVKN

GluCalpha-PA LHLPRFTLEKFLTDYCNSKTNTG--TCLKVDLLFKREFSY
GluCalpha-PC LHLPRFTLEKFLTDYCNSKTNTGEYSCLKVDLLFKREFSY
```

Figure 4.34: The sequence alignment of the mapped region of the isoforms Glucaplha-PA and Glucaplha-PC. The splice region is shown in black.

We used the FunMod modelling pipeline to model structural domains for both isoforms using the same template (PDB 5cfb) from the FunFam, and obtained good models for both (GA341 score of 1 and normalised DOPE score of -0.29). The isoform model structures shared a high structural similarity (SSAP Score above 94.6) and sequence similarity (79.1%) (See Figure 4.35 for the structures).

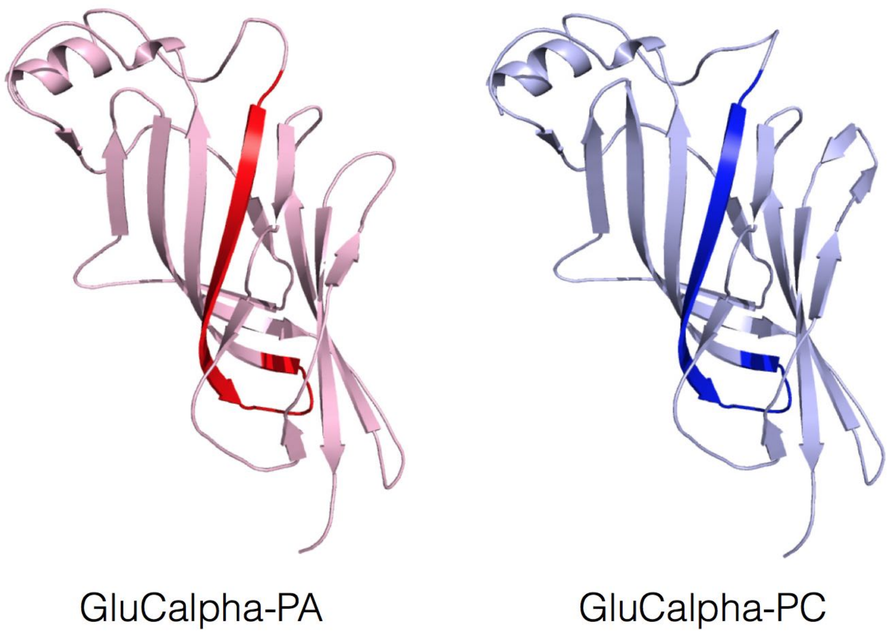
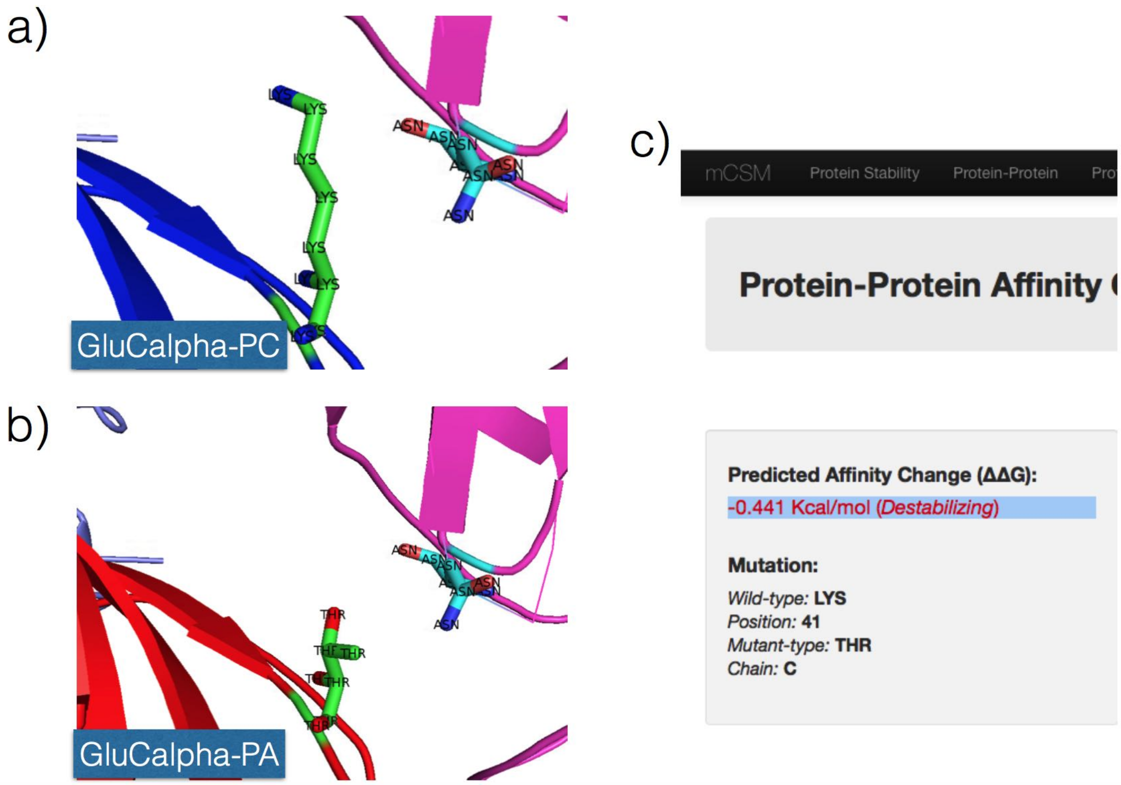


Figure 4.35: The 3D structural model of protein Glucaplha-PA and Glucaplha-PC. The splice regions are colored in a darker shade.

We mapped residues from each MXE exon onto the structural representative. We inspected changes in the residues in the MXE region and found a residue change at position 41 as Glucaplha-PC has a lysine residue whilst Glucaplha-PA has a threonine residue at this position. We used mCSM, a protein-protein affinity mutation predictor (Pires et al., 2014), to find out the effect of this change. The program predicted the residue change to be destabilising (See Figure 4.36).



Predicted Affinity Change (ΔΔG):
-0.441 Kcal/mol *(Destabilizing)*

Mutation:
Wild-type: LYS
Position: 41
Mutant-type: THR
Chain: C

Figure 4.36: Different residue usage at a protein interaction site: lysine residue in Glucaplha-PC versus threonine residue in Glucaplha-PA. The program mCSM (Pires et al., 2014) predicted this change to have destabilising effect on protein-protein affinity. (a) Glucaplha-PC has lysine residue at position 41 (b) Glucaplha-PA has threonine residue at position 41 (c) The mCSM prediction result

4.4.6.3 Pyruvate kinase (PKM-004 (ENSP00000334983) versus PKM-008 (ENSP00000457253))

The PKM gene codes for a pyruvate kinase. Both the isoforms were mapped to 4 FunFams. The sequences for the first three domains are identical. The MXE event we were interested is in the fourth domain. Besides the splicing region, PKM-004 has an additional exon spliced into this domain (see Figures 4.37 and 4.38 for more details). Overall, there are considerable residue changes in the MXE region (see Figure 4.39.)

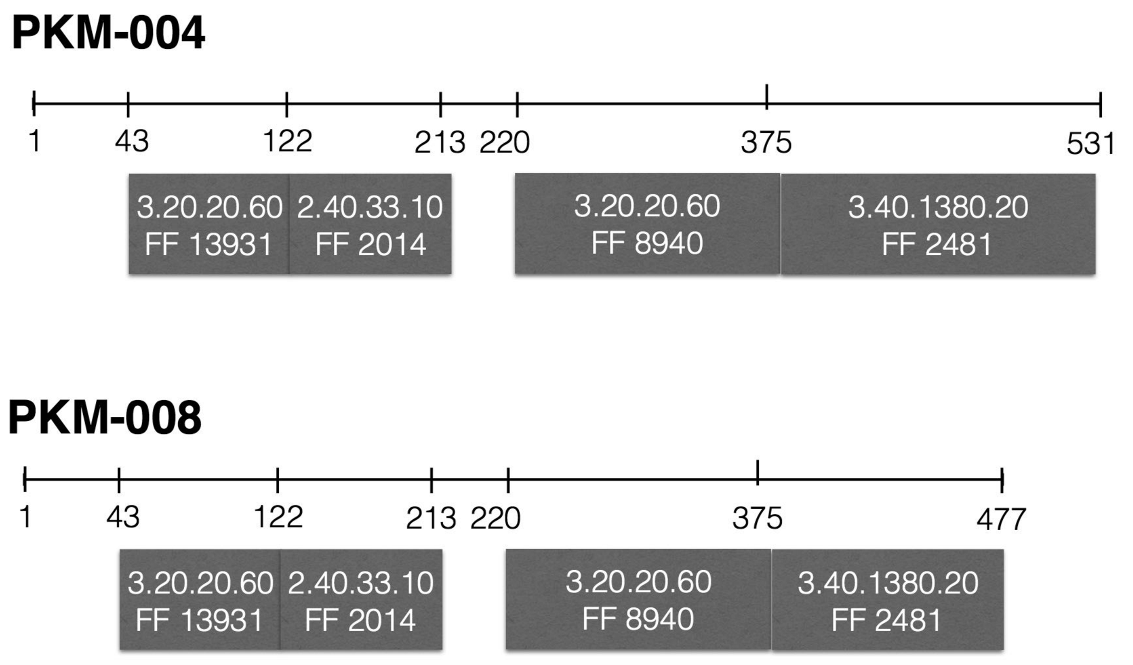


Figure 4.37: Protein structures of the PKM isoforms. The MXE event we were interested is in the fourth domain involving FunFam 3.40.1380.20.FF.2481.

PKM-004

Domain 4 (predicted to belong to 3.40.1380.20.FF2481):

RMQHLIAREAEAAIYHLQLFEELR
RLAPITSDPTEATAVGAVEASFKC
CSGAIIVLTKSGRSAHQVARYRPR
APIIAVTRNPQTARQAHLYRGIFPV
LCKDPVQEAWAEDVDLRVNFAM
NVGKARGFFKKGDVVIVLTGWRP
GSGFTNTMRVVPVP

PKM-008

Domain 4 (predicted to belong to 3.40.1380.20.FF2481):

RMQHLIAREAEAAMFHRKLFEEL
VRASSHSTDLMEAMAMGSVEAS
YKCLAAALIVLTESGRSAHQVARY
RPRAPIIAVTRNPQTARQAHLYRG
IFPVLCKDPVQEAWAED

Figure 4.38: Sequences from domain 4 for the 2 isoforms. The splice regions are colored orange. PKM-004 contains 46 additional residues, colored light blue.

```
PKM-004   IAREAEAAIYHLQLFEELRRLAPITSDPTEATAVGAVEASFKCCSGAIIVLTKSGR
PKM-008   IAREAEAAMFHRKLFEELVRASSHSTDLMEAMAMGSVEASYKCLAAALIVLTESGR
```

Figure 4.39: The sequence alignment of the PKM MXE region. Variable residues are colored red.

There is a known structure available for PKM-004 (PDB ID 1t5a) and domain 4 corresponds to CATH domain 1t5aA01. However, there is no known structure available for PKM-008. We generated a good model for this isoform using the FunMod platform, with GA341 score of 1 and normalised score of -0.87 using 1t5aA01 as the template. Overall, the sequence similarity between isoforms is high (77%) and the SSAP score is very good, 81. Figure 4.40 illustrates the superposition.

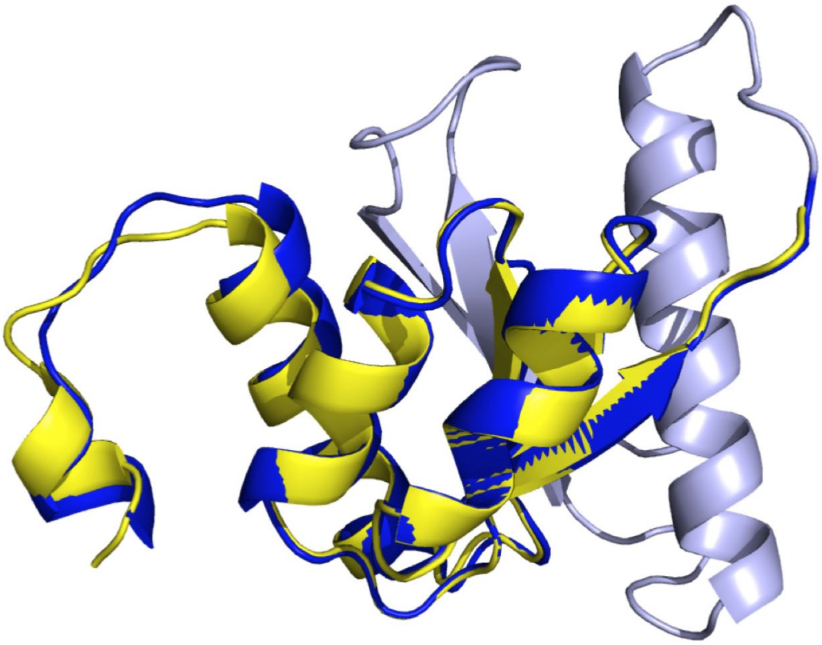


Figure 4.40: Structural superposition of domain 4 of the two isoforms. The blue structure is PKM-004, the yellow is PKM-008. The extra exon sequence in PKM-004 is colored light blue.

We inherited the functional residues of PKM-004 onto PKM-008 to compare the isoforms. There is a ligand (beta-fructose-1,6-diphosphate) binding site located close to the MXE region. Close to this binding site there is a change of residue at position 433, a ligand binding residue. In addition, the ligand binding site also lies close to the extra exon found only in PKM-004. Therefore, we speculated that the binding of the ligand could be modified in isoform PKM-008 (See Figure 4.41).

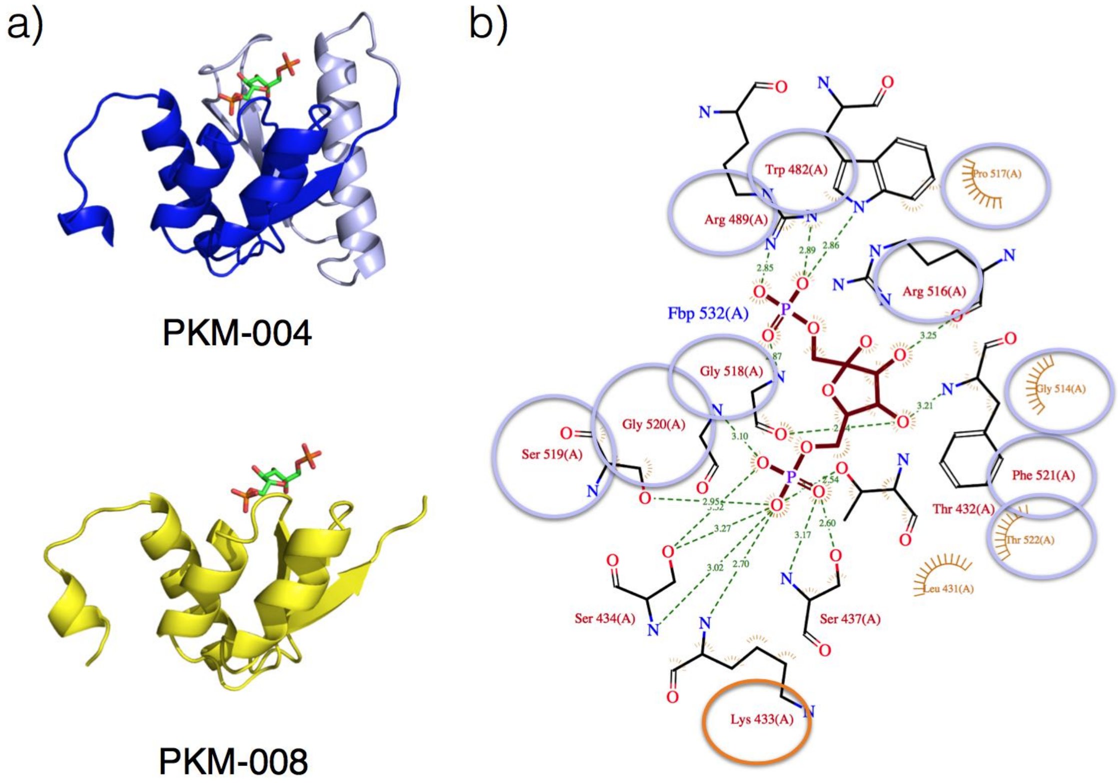


Figure 4.41: The additional exon in PKM-004 (residues colored in light blue in Figure (a)) is involved in the ligand binding with beta-fructose-1,6-diphosphate. These residues are circled in light blue circles also in the ligand-binding plot from PDBsum (De Beer et al., 2013) shown in (b). The variable MXE residues is circled in orange in the ligand-binding plot shown in (b).

In addition, the variable residues within the MXE region may disrupt protein-complex formation. PKM has been demonstrated to oligomerise as a tetramer (Dombrauckas et al., 2005). To form the tetramer, domain 4 of the PKM monomers interact. There are many variable residues located near this interface (15 residues, See Figure 4.42). We used mCSM (Pires et al., 2014) to assess the effect of the MXE variable residue changes on protein-protein affinity. 13 out of the 15 variable residue changes were found to be destabilising to the complex (See Figure 4.43). We therefore speculated that the interactions between these domain 4 regions of the protein may be disrupted, and that therefore, PKM-008 may only form a dimer (See 4.44).

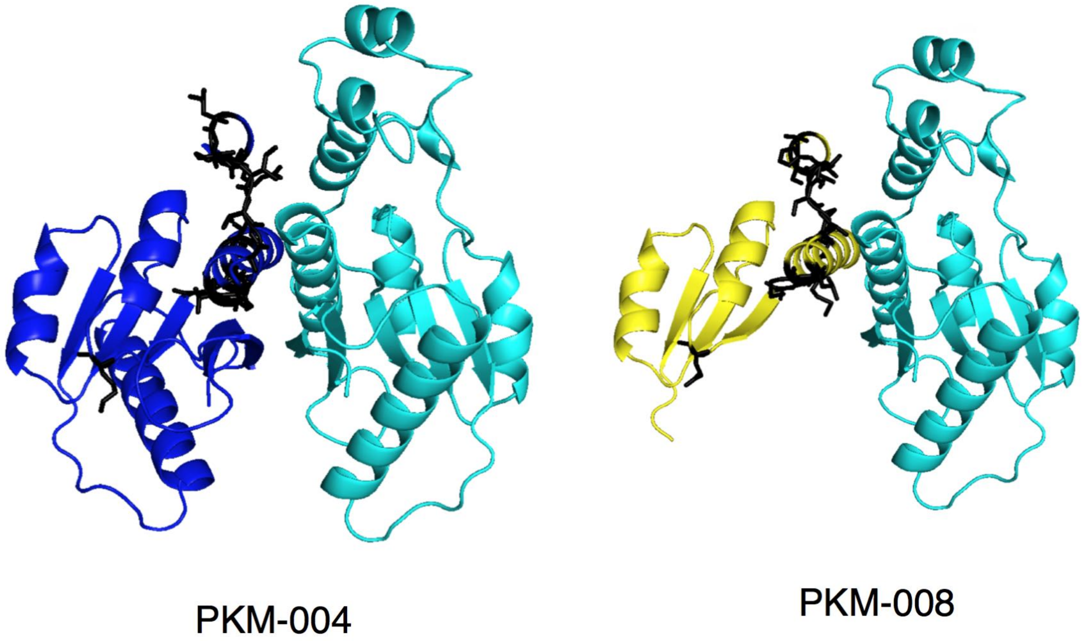
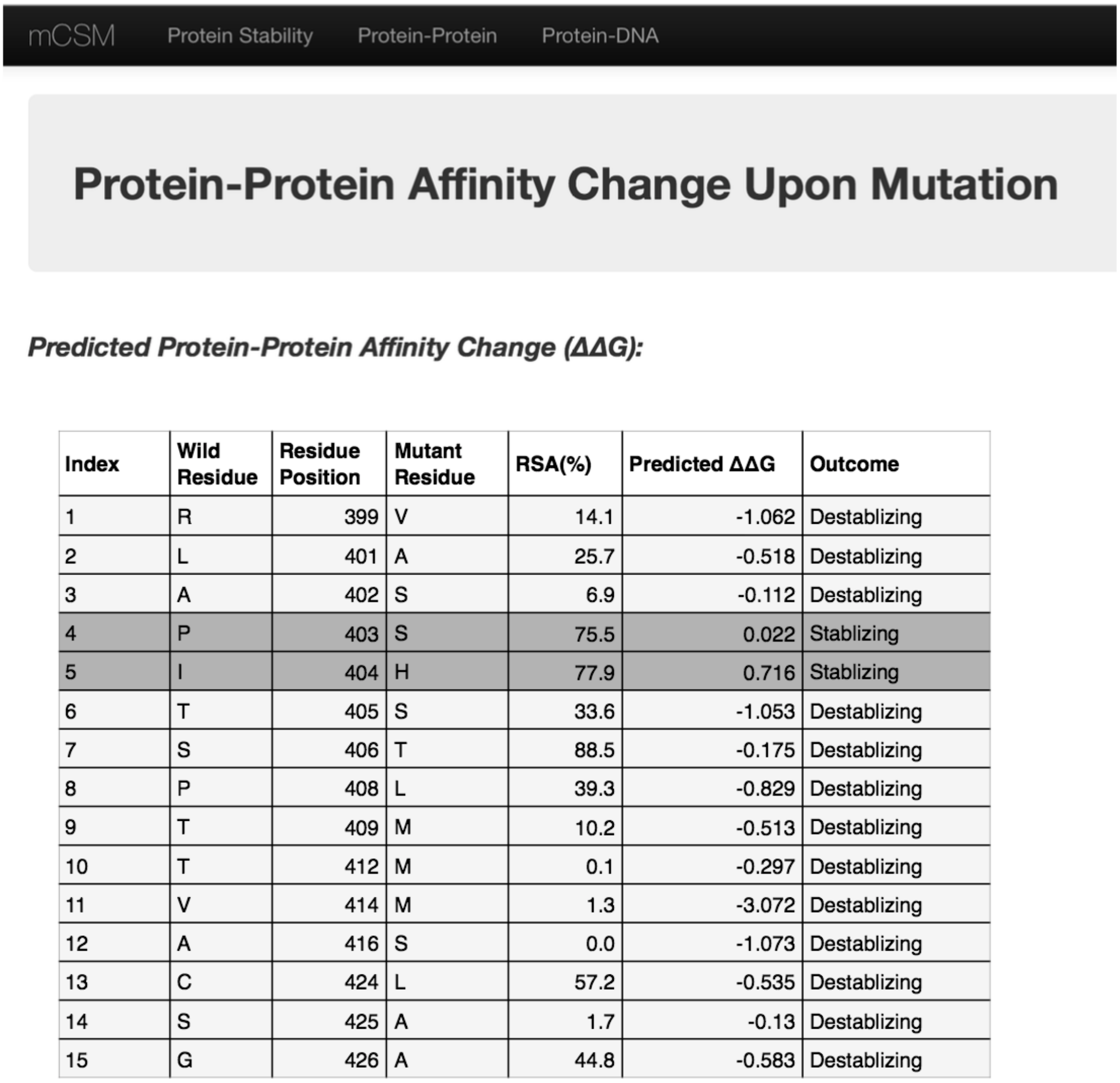


Figure 4.42: Illustration of the two PKM isoforms with MXE variable residues shown in black.



Protein-Protein Affinity Change Upon Mutation

Predicted Protein-Protein Affinity Change (ΔΔG):

Index	Wild Residue	Residue Position	Mutant Residue	RSA(%)	Predicted ΔΔG	Outcome
1	R	399	V	14.1	-1.062	Destablizing
2	L	401	A	25.7	-0.518	Destablizing
3	A	402	S	6.9	-0.112	Destablizing
4	P	403	S	75.5	0.022	Stablizing
5	I	404	H	77.9	0.716	Stablizing
6	T	405	S	33.6	-1.053	Destablizing
7	S	406	T	88.5	-0.175	Destablizing
8	P	408	L	39.3	-0.829	Destablizing
9	T	409	M	10.2	-0.513	Destablizing
10	T	412	M	0.1	-0.297	Destablizing
11	V	414	M	1.3	-3.072	Destablizing
12	A	416	S	0.0	-1.073	Destablizing
13	C	424	L	57.2	-0.535	Destablizing
14	S	425	A	1.7	-0.13	Destablizing
15	G	426	A	44.8	-0.583	Destablizing

Figure 4.43: Summary of the mCSM predicted changes in stability for MXE residue changes located near the protein interface in PKM-008.

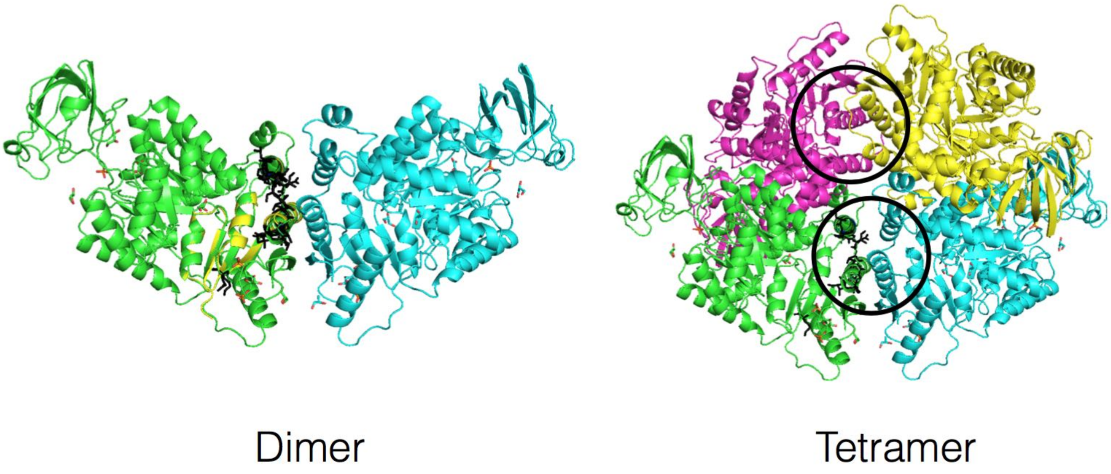


Figure 4.44: The dimeric and tetrameric form of PKM showing MXE variable residues in the dimer interface (on the left). The right hand picture shows the location of the dimer interfaces in the tetramer. These interactions are likely to be affected by the changes in the MXE variable residues. This region is highlighted by a black circle in the figure on the right for PKM-008.

Most somatic cells produce energy predominantly through oxidative phosphorylation, whereas cancer cells produce energy mainly through the less efficient glycolytic pathway, followed by lactic acid production under aerobic conditions (Warburg, 1956). PKM-004 (commonly known as PKM2) has been shown to contribute to cancer metabolism (Wong et al., 2015). PKM2 dimers and tetramers possess low and high levels of enzyme activity, respectively. The PKM2 dimer reduces the flux of glycolysis, therefore redirects all the glucose-derived carbon towards biosynthesis of cellular building blocks, and the conversion of of pyruvate to lactate. On the other hand, the PKM2 tetramer promotes the flux of glycolysis and the massive productive of ATP via a respiratory chain (See Figure 4.45.)

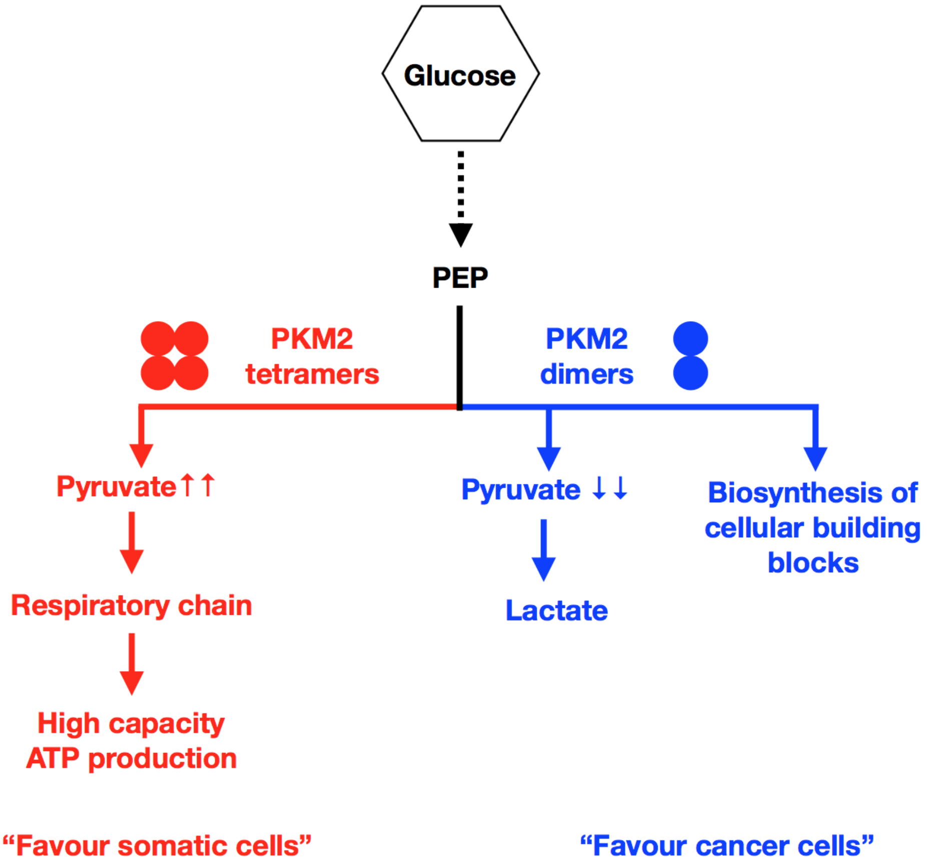


Figure 4.45: Schematic illustration of the role of PKM2 in cancer metabolism

Our analyse suggest that PKM-008 that is unlikely to form a tetramer. We mapped the cancer associated mutations from the COSMIC database to both PKM-004 and PKM-008. The COSMIC residues coincide with variable residues in the interface of both isoforms. Currently, no experimental studies have been done on PKM-008 to examine involvement in cancer metabolism. However, our analysis suggests that there may be a role.

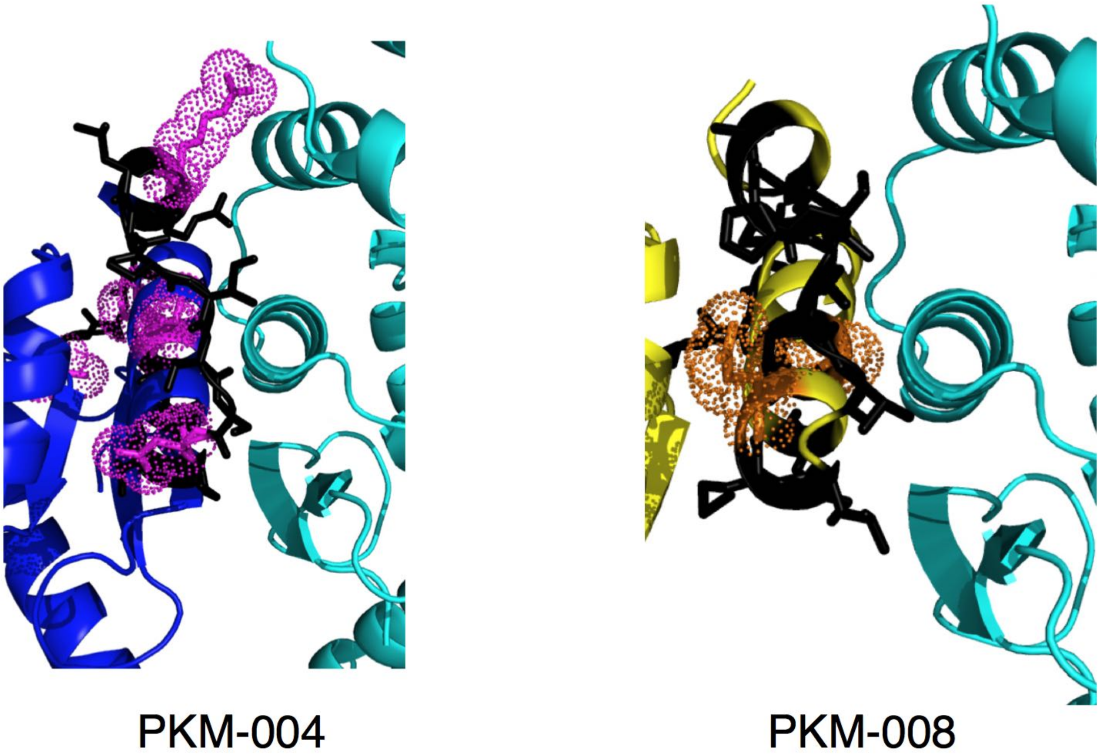


Figure 4.46: COSMIC mutation residues lie close to the interface of both PKM isoforms. MXE variable residues were colored in black, COSMIC mutation residues were colored in purple/orange.

4.5 Conclusions

Mutually exclusive exons are characterised by coordinated splicing of exons such that only one of the two exons is retained, while the other is spliced out. These events are enriched in proteomics data, suggesting a functional role. Previous analysis by other groups have demonstrated that the MXEs are evolutionary conserved among vertebrates (Abascal et al., 2015b) and also demonstrated that these MXE regions are highly sequence similar and that MXE events do not usually disrupt the structural domain. However, to date, there are no reported studies in the literature on the likely functional consequences of MXE events, using structural models.

In this chapter, we first identified the MXE events in human and fly datasets using an in-house prediction protocol. Most of these MXE genes are associated with membrane proteins (associated with cell-cell adhesion, signal transduction and molecule/ion transport) and enzymes (e.g. kinases, transferases, hydrolases and oxidoreductases). We managed to annotated about 50% of human and fly MXE with structural information (i.e. known structures, structural models built using the FunMod modelling platform described in Chapter 2 and structural mapping) and demonstrated that about 15% of the MXEs are disordered.

For those MXE events which can be mapped to a structure, they usually have less than 10 variable residues and they are more exposed to the solvent than the whole spliced region. Variable residues usually exhibit a significant changes in their physio-chemical property.We also demonstrated these variable residues have a tendency to lie close to a known or predicted functional site. These results suggest MXE events may have a number of important roles in cells generally and also in cancer.

Chapter 5

Conclusions

The last few years have been an exciting era for the protein structural modelling community. There have been substantial improvements in residue-contact prediction thanks to the use of direct coupling analysis (Jones et al., 2012; Kamisetty et al., 2013; Marks et al., 2011; Nugent and Jones, 2012), better statistical machine learning (Adhikari and Cheng, 2016; Eickholt and Cheng, 2012; Feinauer et al., 2014) and the huge amount of new sequence data that is being provided by metagenome analyses (Ovchinnikov et al., 2017). Many groups are now employing residue-contact prediction to enhance the performance of their methods (Adhikari and Cheng, 2016; Park et al., 2016). Better profile methods such as conditional random forest (Joo et al., 2016) and Markov random fields (Ma et al., 2014) have improved the accuracy of the template-selection process. There have also been improvements in model quality assessment methods (through the use of integrated approaches and promising new approaches using deep learning) (Cao et al., 2016; Jing et al., 2016; Uziela et al., 2017, 2016) and the structural refinement category (with improved energy functions and MD simulations) (Della Corte et al., 2016; Feig, 2016; Kim and Kihara, 2016; Lee et al., 2016; Park et al., 2016).

In the first work chapter, we demonstrated the value of organizing domain superfamilies into functional families (FunFams) for template selection in modelling protein monomers. FunFams group relatives that are highly likely to be of similar structure and function. They are generated using a new functional sub-classification protocol named FunFHMMer (Das et al., 2015) which constrains clustering of relatives by ensuring that any new relatives joining a particular cluster match the highly conserved functional determinants for that cluster (for example likely specificity-determining residues that influence the type of compounds bound or protein interactions).

The improvement in accuracy for template selection relative to the HMM-based strategy used by HHsearch is therefore likely to be owing to the fact that the Fun-

Fam template-selection process only allows very remote relatives to be selected if they share the same or highly similar residues at key functional sites. Although HHsearch uses a powerful search strategy for remote homologues, there is no explicit constraint to ensure those equivalent functional residues are matched. Preliminary results demonstrated that FunFams models (modelled using remote templates) have a higher local agreement with the native structure compared to HHsearch models. Having a local similarity is important when we are modelling enzymes or protein complexes, where a better representation of functional sites is crucial. Further investigation could be done to confirm this finding.

We developed a structural modelling platform named FunMoD (Lam et al., 2017, 2016) that utilised FunFams to select the structural template. Since FunFams only cover about 70% of the structural domains in CATH, we employed also the HHsearch strategy that uses different structural template libraries to increase the structural coverage of human and fly proteomes. This modelling platform has been applied to model up to 70% of structurally uncharacterised sequences in human and fly genomes. In addition, our analysis demonstrated the advantages of different template libraries (i.e. there are many query sequences for which only one strategy can build a good quality model).

In our study, we tested a few model quality assessment methods and found that all of them performed similarly in selecting the best model. This result conflicts with results obtained by other research groups. In theory, recent advanced methods such as ProQ2 (Ray et al., 2012) and ModFOLDclust2 (McGuffin and Roche, 2010) (that have been performed well in the recent CASP model quality assessment category) should outperform classical statistical potential methods such as nDOPE (Shen and Sali, 2006). This agreement between the methods may be due to the low number of models built for each target (i.e. It is easy to select the best model from 10 decoys) or the benchmark target dataset used (that consists of mainly close homologue query targets). Further studies could be carried out to investigate this issue.

We also tested two different alignment methods (HHsearch and MAFFT) and found the HHsearch alignment method to be better. We speculate this this is due to the maximum alignment algorithm employed by HHsearch (that maximises the number of correctly aligned pairs of residues) and also the filtering step that removes sequences that are too distant. In order to further improve our template and sequence alignment protocol, newer/better sequence alignment methods such as conditional random forest and Markov random field can be explored. Also, the inclusion of additional templates may improve the model quality, particularly by extending the coverage of the query sequence (Larsson et al., 2008) or when the templates have structurally complementary (Chakravarty et al., 2008). Multiple templates also provide conserved distance constraints, which are not available to single-template protocols (Cheng, 2008).

Future work could explore the use of better loop/side-chain modelling protocols (i.e. SCWRL4 (Krivov et al., 2009)), newer structural refinement protocols that involve running MD simulations with physics-based force fields (e.g. AMBER ff14SB (Maier et al., 2015), and CHARMM36m (Huang et al., 2017)), hybrid potentials which employ both knowledge-based and physics-based energy functions, building missing parts in the template regions using *ab initio* methods, and possibly using co-evolution information in the structural modelling platform.

In the second work chapter, we demonstrated the value of FunFams in protein complex modelling. Domain-domain interfaces in FunFams tend to be more conserved in FunFams than across the CATH superfamily. Intra-chain interfaces are more conserved than inter-chain interfaces. The difference between conservation within FunFams and across a superfamily is particularly pronounced for inter-chain interfaces especially at low sequence identity. Previously, Aloy et al. demonstrated that pairs of proteins with sequence identity >40% tend to interact similarly. In our analysis, we found a lot of domain-domain interfaces in FunFams to be conserved even when sequence identity is lower than 40%. This is a very encouraging result and suggests that it is capable to use FunFams in protein complex modelling.

We used FunFams in two different modes of template based complex modelling methods, one based on sequence similarity and one on structural similarity. We demonstrated the value of using FunFams to better identify structural templates for the sequence based methods. Our FunFam based protocol gave significantly more good quality models compared to a BLAST based protocol and slightly better than an HHsearch based protocol. This again suggests that it may be valuable to use a functional family-based protocol to guide template selection in binary protein complex modelling.

The quality of models of the structural similarity based modelling methods is highly dependent on the template library. We devised a new template library using FunFams to remove the remove the redundancy. The template library gave 10% more good quality models than the widely used publicly available DOCKGROUND template library. Our protocol managed to rank 44% of good quality models in the Top10. This is an encouraging start. It is difficult to compare this result against the results of other groups, since they do not report these statistics. Further work could be carried out to improve the protocol by incorporating methods employing energy based functions (i.e. FiberDOCK (Mashiach et al., 2010)) or developing a machine learning approach based on different molecular distributors. Clearly this work would need to subject to independent review by participation in CAPRI.

However, we found some domain-domain interactions within FunFams have low conservation. Possible reasons are due to the pollution of the FunFams (FunFams are polluted with sequences that are annotated with diverse GO terms), different multi-domain architectures (MDAs), different oligomeric states, crystal contacts and other possible issues. Ongoing research in the Orengo group is involved with subclassifying these polluted FunFams into functionally pure clusters.

Although there has been significant improvement over the years in complex modelling, there are still very challenging cases that involve multiple binding patches, cases that involve small interface regions or cases where interface residues lie in loop regions (Lensink et al., 2017; Soni and Madhusudhan, 2017). Further research is clearly still needed to improve the complex modelling.

In the final work chapter, we employed our structural modelling tool to understand the implications of alternative splicing. Alternative splicing (AS) has been suggested as one of the major processes expanding the diversity of proteomes in multicellular organisms. We focused on mutually exclusively exons (MXEs) for which there is more evidence of translation into proteins from proteomic experiments. Previous analyses by other groups have demonstrated that the MXEs are evolutionary conserved among vertebrates (Abascal et al., 2015b) and also demonstrated that these MXE regions are highly sequence similar and that MXE events do not usually disrupt the structural domain (Tress et al., 2017). However, to date, there are no reported large-scale studies in the literature on the likely functional consequences of MXE events, using structural analysis.

We first identified the MXE events in human and fly datasets using an in-house prediction protocol. Most of these MXE genes are associated with membrane proteins (associated with cell-cell adhesion, signal transduction, and molecule/ion transport) and enzymes (e.g. kinases, transferases, hydrolases, and oxidoreductases). We managed to annotate about 50% of human and fly MXE with structural information (i.e. known structures, structural models built using the FunMod modelling platform and structural mapping) and demonstrated that about 15% of the MXEs are disordered.

For those MXE events which can be mapped to a structure, they usually have less than 10 variable residues and they are more exposed to the solvent than the whole spliced region. Variable residues usually exhibit significant changes in their physiochemical property. We also demonstrated these variable residues have a tendency to lie close to a known or predicted functional site. Coupled with the fact that MXE splicing tends to be preserved at the proteomics level, our results suggest MXE events may have a number of important roles in cells generally and also in cancer.

References

Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J. M., del Pozo, A., Vázquez, J., Valencia, A., and Tress, M. L. (2015a). Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level. *PLoS Comput Biol*, 11(6):e1004325. 155, 156

Abascal, F., Tress, M. L., and Valencia, A. (2015b). The evolutionary fate of alternatively spliced homologous exons after gene duplication. *Genome biology and evolution*, 7(6):1392–1403. 157, 201, 206

Abhiman, S. and Sonnhammer, E. L. L. (2005). FunShift: a database of function shift analysis on protein subfamilies. *Nucleic acids research*, 33(Suppl 1):D197–D200. 49

Adhikari, B. and Cheng, J. (2016). Protein Residue Contacts and Prediction Methods. *Methods in molecular biology (Clifton, N.J.)*, 1415:463–476. 56, 57, 202

Ahnert, S. E., Marsh, J. A., Hernández, H., Robinson, C. V., and Teichmann, S. A. (2015). Principles of assembly reveal a periodic table of protein complexes. *Science*, 350(6266):aaa2245. 94

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., et al. (2016). The ensembl gene annotation system. *Database*, 2016:baw093. 48

Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., et al. (2014). The Structure–Function Linkage Database. *Nucleic Acids Research*, 42(D1):D521–D530. 50, 66

Alekseyenko, A. V., Kim, N., and Lee, C. J. (2007). Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *Rna*, 13(5):661–670. 149

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., and Burgess, E. (2005). The biomolecular interaction network database and related tools 2005 update. *Nucleic acids research*, 33(Suppl 1):D418–D424. 105

Allen, G. S. and Stokes, D. L. (2013). Modeling, Docking, and Fitting of Atomic Structures to 3D Maps from Cryo-Electron Microscopy. In *Methods in molecular biology (Clifton, N.J.)*, volume 955, pages 229–241. 89

Aloy, P., Ceulemans, H., Stark, A., and Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *Journal of molecular biology*, 332(5):989–998. 101, 109, 204

Aloy, P. and Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nature biotechnology*, 22(10):1317–1321. 91

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410. 36, 53

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402. 36, 53

Andreani, J., Faure, G., and Guerois, R. (2013). Interevscore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics*, 29(14):1742–1749. 146

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2014). SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Research*, 42(D1):D310–D314. 39, 46, 50, 51

Ansari, S. and Helms, V. (2005). Statistical analysis of predominantly transient protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 61(2):344–355. 92

Arafat, H., Lazar, M., Salem, K., Chipitsyna, G., Gong, Q., Pan, T.-C., Zhang, R.-Z., Yeo, C. J., and Chu, M.-L. (2011). Tumor-specific expression and alternative splicing of the COL6A3 gene in pancreatic cancer. *Surgery*, 150(2):306–315. 154

Bahadur, R. P., Chakrabarti, P., Rodier, F., and Janin, J. (2004). A dissection of specific and non-specific protein–protein interfaces. *Journal of molecular biology*, 336(4):943–955. 92

Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, 294(5540):93–96. 53

Baker, M. L., Chen, M., Durmaz, T., Baldwin, P., Ju, T., and Ludtke, S. J. (2016). Building and Validating Atomic Models for Cryo-EM Density Maps. *Microscopy and Microanalysis*, 22(Suppl 3):2080–2081. 89

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041. 188

Bartesaghi, A., Merk, A., Banerjee, S., Matthies, D., Wu, X., Milne, J. L., and Subramaniam, S. (2015). 2.2 å resolution cryo-em structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science*, 348(6239):1147–1151. 27

Bates, P. A., Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. E. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins: Structure, Function, and Bioinformatics*, 45(Suppl 5):39–46. 55

Benkert, P., Künzli, M., and Schwede, T. (2009). QMEAN server for protein model quality estimation. *Nucleic acids research*, 37(Suppl 2):W510–W514. 63

Benkert, P., Tosatto, S. C. E., and Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics*, 71(1):261–277. 59, 60, 63

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, 28(1):235–242. 24, 36, 52, 95

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic acids research*, 42(W1):W252–W258. 38

Bienert, S., Waterhouse, A., de Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research*, 45(D1):D313–D319. 62, 63

Branden, C. I. and Tooze, J. (1999). *Introduction to protein structure*. Garland Science. 58

Brasch, J., Harrison, O. J., Honig, B., and Shapiro, L. (2012). Thinking outside the cell: how cadherins drive adhesion. *Trends in cell biology*, 22(6):299–310. 152

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, 4(2):187–217. 55, 58, 59

Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O., Yu, L., Wright, J., Verstraten, R., Adams, D. J., Harrow, J., Choudhary, J. S., et al. (2011). Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and “resurrected” pseudogenes in the mouse genome. *Genome research*, 21(5):756–767. 155

Buchan, D. W. and Jones, D. T. (2017). EigenTHREADER: Analogous protein fold recognition by efficient contact map threading. *Bioinformatics (Oxford, England)*, 33:2684–2690. 65

Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K., and Jones, D. T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic acids research*, 41(W1):W349–W357. 38, 64

Bugge, K., Papaleo, E., Haxholm, G. W., Hopper, J. T. S., Robinson, C. V., Olsen, J. G., Lindorff-Larsen, K., and Kragelund, B. B. (2016). A combined computational and structural model of the full-length human prolactin receptor. *Nature communications*, 7:11578. 90

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., and Babu, M. M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular cell*, 46(6):871–883. 152, 153

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10(1):421. 50, 112

Cao, R., Adhikari, B., Bhattacharya, D., Sun, M., Hou, J., and Cheng, J. (2016). QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics*, 33(4):586–588. 59, 61, 202

Cao, R., Bhattacharya, D., Adhikari, B., Li, J., and Cheng, J. (2015). Large-scale model quality assessment for improving protein tertiary structure prediction. *Bioinformatics*, 31(12):i116—-i123. 53

Cao, R., Jo, T., and Cheng, J. (2016). Evaluation of Protein Structural Models Using Random Forests. *ArXiv e-prints*. 59

Cao, R., Wang, Z., and Cheng, J. (2014). Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment. *BMC structural biology*, 14(1):13. 61

Chae, M., Krull, F., and Knapp, E. (2015). Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 83(5):881–890. 59

Chakrabarti, P. and Janin, J. (2002). Dissecting protein–protein recognition sites. *Proteins: Structure, Function, and Bioinformatics*, 47(3):334–343. 92

Chakravarty, S., Godbole, S., Zhang, B., Berger, S., and Sanchez, R. (2008). Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. *BMC Struct Biol*, 8:31. 204

Chandonia, J., Hon, G., Walker, N. S., Conte, L. L., Koehl, P., Levitt, M., and Brenner, S. E. (2004). The ASTRAL compendium in 2004. *Nucleic acids research*, 32(Suppl 1):D189–D192. 94

Chen, M., David, C. J., and Manley, J. L. (2012). Concentration-dependent control of pyruvate kinase M mutually exclusive splicing by hnRNP proteins. *Nature structural & molecular biology*, 19(3):346–354. 156

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2009). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):12–21. 58

Cheng, J. (2008). A multi-template combination algorithm for protein comparative modeling. *BMC structural biology*, 8(2):18. 204

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823. 30

Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J. (2004). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, 20(1):45–50. 100, 101

Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G., and Orengo, C. A. (2015). Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics*, 31(21):3460–3467. 35, 49, 50, 66, 118, 164, 202

David, A. and Sternberg, M. J. E. (2015). The contribution of missense mutations in core and rim residues of protein–protein interfaces to human disease. *Journal of molecular biology*, 427(17):2886–2898. 166, 181

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., and Sillitoe, I. (2017). CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, 45(D1):D289–D295. 39, 45, 46, 65, 159

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, volume 5, pages 345–352. National Biomedical Research Foundation Silver Spring, MD. 31

De Beer, T. A., Berka, K., Thornton, J. M., and Laskowski, R. A. (2013). Pdbsum additions. *Nucleic acids research*, 42(D1):D292–D296. 195

Della Corte, D., Wildberg, A., and Schröder, G. F. (2016). Protein structure refinement with adaptively restrained homologous replicas. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl 1):302–313. 57, 88, 202

Dessailly, B. H., Dawson, N. L., Mizuguchi, K., and Orengo, C. A. (2013). Functional site plasticity in domain superfamilies. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1834(5):874–889. 35, 164

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J. M., Taly, J. F., and Notredame, C. C. . (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res*, 39(Suppl 2):W13–7. 34, 35

Dombrauckas, J. D., Santarsiero, B. D., and Mesecar, A. D. (2005). Structural Basis for Tumor Pyruvate Kinase M2 Allosteric Regulation and Catalysis. *Biochemistry*, 44(27):9417–9429. 195

Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737. 100, 101

Dong, G. Q., Fan, H., Schneidman-Duhovny, D., Webb, B., and Sali, A. (2013). Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics*, 29(24):3158–3166. 59, 112, 146

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of molecular biology*, 347(4):827–839. 48, 161

Douguet, D., Chen, H.-C., Tovchigrechko, A., and Vakser, I. A. (2006). Dockground resource for studying protein–protein interfaces. *Bioinformatics*, 22(21):2612–2618. 93, 117, 118

Dutertre, M., Vagner, S., and Auboeuf, D. (2010). Alternative splicing and breast cancer. *RNA biology*, 7(4):403–411. 153

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology*, 7(10):e1002195. 37, 53, 65, 159

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797. 34, 35

Eickholt, J. and Cheng, J. (2012). Predicting protein residue–residue contacts using deep networks and boosting. *Bioinformatics*, 28(23):3066–3072. 56, 202

Ellis, J. D., Barrios-Rodiles, M., Çolak, R., Irimia, M., Kim, T., Calarco, J. A., Wang, X., Pan, Q., O'Hanlon, D., and Kim, P. M. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular cell*, 46(6):884–892. 152

Eramian, D., Eswar, N., Shen, M., and Sali, A. (2008). How well can the accuracy of comparative protein structure models be predicted? *Protein Science*, 17(11):1881–1893. 62, 63

Esquivel-Rodríguez, J., Yang, Y. D., and Kihara, D. (2012). Multi-lzerd: Multiple protein docking for asymmetric complexes. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1818–1833. 101

Eswar, N., Eramian, D., Webb, B., Shen, M.-Y., and Sali, A. (2008). *Protein structure modeling with MODELLER*, pages 145–159. Springer. 82, 144

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., Pieper, U., Stuart, A. C., Marti-Renom, M. A., Madhusudhan, M. S., and Yerkovich, B. (2003). Tools for comparative protein structure modeling and analysis. *Nucleic acids research*, 31(13):3375–3380. 62

Ezkurdia, I., del Pozo, A., Frankish, A., Rodriguez, J. M., Harrow, J., Ashman, K., Valencia, A., and Tress, M. L. (2012). Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Molecular biology and evolution*, 29(9):2265–2283. 155

Feig, M. (2016). Local Protein Structure Refinement via Molecular Dynamics Simulations with locPREFMD. *Journal of Chemical Information and Modeling*, 56(7):1304–1312. 54, 57, 88, 202

Feinauer, C., Skwark, M. J., Pagnani, A., and Aurell, E. (2014). Improving Contact Prediction along Three Dimensions. *PLoS Computational Biology*, 10(10). 56, 202

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., and Fraser, M. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic acids research*, 45(D1):D190–D199. 63

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–D285. 37, 47, 50

Finn, R. D., Miller, B. L., Clements, J., and Bateman, A. (2014). iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Research*, 42(D1):D364–D373. 95, 97, 108

Fiser, A. (2004). Protein structure modeling in the proteomics era. *Expert review of proteomics*, 1(1):97–110. 53

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic acids research*, 45(D1):D777–D783. 165

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2013). Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309. 51

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152. 67

Furnham, N., Holliday, G. L., De Beer, T. A. P., Jacobsen, J. O. B., Pearson, W. R., and Thornton, J. M. (2014). The Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research*, 42(D1):D485–D489. 163

Gabellini, D., D'antona, G., Moggio, M., Prelle, A., Zecca, C., Adami, R., Angeletti, B., Ciscato, P., Pellegrino, M. A., Bottinelli, R., et al. (2006). Facioscapulohumeral muscular dystrophy in mice overexpressing frg1. *Nature*, 439(7079):973. 153

Gao, M. and Skolnick, J. (2010). Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences*, 107(52):22517–22522. 91

Gardina, P. J., Clark, T. A., Shimada, B., Staples, M. K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., et al. (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC genomics*, 7(1):325. 154

Ghoorah, A. W., Devignes, M.-D., Smail-Tabbone, M., and Ritchie, D. W. (2014). KBDOCK 2013: a spatial classification of 3D protein domain family interactions. *Nucleic acids research*, 42(D1):D389–D395. 97

Goncearenco, A., Shoemaker, B. A., Zhang, D., Sarychev, A., and Panchenko, A. R. (2014). Coverage of protein domain families with structural protein–protein interactions: current progress and future trends. *Progress in biophysics and molecular biology*, 116(2):187–193. 91

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., et al. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, 471(7339):473–479. 150

Gupta, S., Barrett, T., Whitmarsh, A. J., Cavanagh, J., Sluss, H. K., Derijard, B., and Davis, R. J. (1996). Selective interaction of jnk protein kinase isoforms with transcription factors. *The EMBO journal*, 15(11):2760. 157

Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., and Schwede, T. (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database*, 2013:bat031. 63

Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J., and Orengo, C. (2003). Recognizing the fold of a protein structure. *Bioinformatics*, 19(14):1748–1759. 41

Hatje, K. and Kollmar, M. (2014). Kassiopeia: a database and web application for the analysis of mutually exclusive exomes of eukaryotes. *BMC genomics*, 15(1):115. 156

He, B., Mortuza, S. M., Wang, Y., Shen, H.-B., and Zhang, Y. (2017). NeBcon: Protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics (Oxford, England)*, 33(15):2296–2306. 56

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919. 31

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233(1):123–138. 39, 43, 50

Holm, L. and Sander, C. (1994). Parser for protein folding units. *Proteins: Structure, Function, and Bioinformatics*, 19(3):256–268. 47

Holmes, I. and Durbin, R. (1998). Dynamic programming alignment accuracy. *Journal of computational biology*, 5(3):493–504. 38

Hu, Z., Scott, H. S., Qin, G., Zheng, G., Chu, X., Xie, L., Adelson, D. L., Oftedal, B. E., Venugopal, P., Babic, M., et al. (2015). Revealing missing human protein isoforms based on Ab initio prediction, RNA-seq and proteomics. *Scientific reports*, 5. 150

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57. 159

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., and MacKerell Jr, A. D. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14:71–73. 54, 204

Huang, Y.-F. (2007). *Study of Mining Protein Structural Properties and its Application*. PhD thesis, National Taiwan University. 68

Hubbard, S. J. and Thornton, J. M. (1993). Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*. 118, 162

Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). Protein–protein docking benchmark version 4.0. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3111–3114. 105, 120, 142

Im, W., Liang, J., Olson, A., Zhou, H.-X., Vajda, S., and Vakser, I. A. (2016). Challenges in structural approaches to cell modeling. *Journal of molecular biology*, 428(15):2943–2964. 146

Jacob, J., Haspel, J., Kane-Goldsmith, N., and Grumet, M. (2002). L1 mediated homophilic binding and neurite outgrowth are modulated by alternative splicing of exon 2. *Journal of neurobiology*, 51(3):177–189. 154

Janin, J., Bahadur, R. P., and Chakrabarti, P. (2008). Protein–protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(2):133–180. 92

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., Vakser, I., and Wodak, S. J. (2003). CAPRI: a critical assessment of predicted interactions. *Proteins: Structure, Function, and Bioinformatics*, 52(1):2–9. 105

Janin, J., Wodak, S. J., Lensink, M. F., and Velankar, S. (2015). Assessing Structural Predictions of Protein–Protein Recognition: The CAPRI Experiment. *Reviews in Computational Chemistry Volume 28*, pages 137–173. 106

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., et al. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1):184. 50, 66

Jing, X., Wang, K., Lu, R., and Dong, Q. (2016). Sorting protein decoys by machine-learning-to-rank. *Scientific reports*, 6. 59, 61, 202

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202. 64, 88

Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190. 56, 202

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: combining co-evolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):999–1006. 56

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282. 35

Jones, S. and Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. *Progress in biophysics and molecular biology*, 63(1):31–65. 92

Jones, S. and Thornton, J. M. (1996). Principles of protein protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20. 92

Joo, K., Joung, I., Lee, S. Y., Kim, J. Y., Cheng, Q., Manavalan, B., Joung, J. Y., Heo, S., Lee, J., Nam, M., et al. (2016). Template based protein structure modeling by global optimization in CASP11. *Proteins: Structure, Function and Bioinformatics*, 84(Suppl 1):221–232. 38, 57, 202

Jordan, R., Wu, F., Dobbs, D., and Honavar, V. (2011). Protindb: A database of protein-protein interface residues. *Iowa State University (Manuscript in Preparation)*. 95, 107

Jubb, H. C., Pandurangan, A. P., Turner, M. A., Ochoa-Montaño, B., Blundell, T. L., and Ascher, D. B. (2016). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in biophysics and molecular biology*. 166, 181

Kalman, M. and Ben-Tal, N. (2010). Quality assessment of protein model-structures using evolutionary conservation. *Bioinformatics*, 26(10):1299–1307. 59

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39):15674–156749. 56, 202

Kastritis, P. L. and Bonvin, A. M. (2010). Are scoring functions in protein- protein docking ready to predict interactomes? clues from a novel binding affinity benchmark. *Journal of proteome research*, 9(5):2216–2225. 143

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780. 34, 160

Kawabata, T. (2016). HOMCOS: an updated server to search and model complex 3D structures. *Journal of Structural and Functional Genomics*, pages 1–17. 102

Kelley, L. A. (2017). Fold recognition. In *From Protein Structure to Function with Bioinformatics*, pages 59–90. Springer. 65

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, 10(6):845–858. 53, 64, 65

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355. 150

Kihara, D. and Skolnick, J. (2003). The PDB is a covering set of small protein structures. *Journal of molecular biology*, 334(4):793–802. 41

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic acids research*, 35(1):125–131. 149

Kim, H. and Kihara, D. (2016). Protein structure prediction using residue- and fragment-environment potentials in CASP11. *Proteins: Structure, Function and Bioinformatics*, 84(Suppl 1):105–117. 54, 57, 88, 202

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. *Nature*, 509(7502):575–581. 38, 155

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, 2011:bar030. 158

Kleywegt, G. J. and Jones, T. A. (1997). Detecting folding motifs and similarities in protein structures. *Methods in enzymology*, 277:525–545. 43, 64

Koehl, P. and Delarue, M. (1995). A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nature Structural & Molecular Biology*, 2(2):163–170. 55

Kolodny, R., Koehl, P., and Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *Journal of molecular biology*, 346(4):1173–1188. 43

Konopka, B. M., Nebel, J.-C., and Kotulska, M. (2012). Quality assessment of protein model-structures based on structural and functional similarities. *BMC bioinformatics*, 13(1):242. 60

Kopp, J. and Schwede, T. (2004). The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic acids research*, 32(Suppl 1):D230—-D234. 55

Kosciolek, T. and Jones, D. T. (2016). Accurate contact predictions using covariation techniques and machine learning. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl 1):145–51. 56

Krissinel, E. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, 372(3):774–797. 93

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function and Bioinformatics*, 77(4):778–795. 54, 65, 204

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994). Hidden markov models in computational biology: Applications to protein modeling. *Journal of molecular biology*, 235(5):1501–1531. 37

Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T., and Tramontano, A. (2015). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins: Structure, Function, and Bioinformatics*, 82(Suppl 1):349–369. 71

Kundrotas, P. J. and Vakser, I. A. (2013). Global and local structural similarity in protein protein complexes: Implications for template based docking. *Proteins: Structure, Function, and Bioinformatics*, 81(12):2137–2142. 103, 114, 120, 140

Kundrotas, P. J., Zhu, Z., Janin, J., and Vakser, I. A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. *Proceedings of the National Academy of Sciences*, 109(24):9438–9441. 102, 105

Kundrotas, P. J., Zhu, Z., and Vakser, I. A. (2010). GWIDD: Genome-wide protein docking database. *Nucleic Acids Research*, 38(Suppl 1):D513–D517. 97

Kuzu, G., Gursoy, A., Nussinov, R., and Keskin, O. (2013). Exploiting conformational ensembles in modeling protein–protein interactions on the proteome scale. *Journal of proteome research*, 12(6):2641–2653. 103, 118

Lam, S., Das, S., Sillitoe, I., and Orengo, C. (2017). An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences. *Acta Crystallographica Section D: Structural Biology*, 73(8):628–640. 203

Lam, S. D., Dawson, N. L., Das, S., Sillitoe, I., Ashford, P., Lee, D., Lehtinen, S., Orengo, C. A., and Lees, J. G. (2016). Gene3D: expanding the utility of domain assignments. *Nucleic Acids Research*, 44(D1):D404–D409. 37, 48, 64, 65, 203

Lareau, L. F. and Brenner, S. E. (2015). Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Molecular biology and evolution*, 32(4):1072–1079. 155

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., et al. (2007). Clustal w and clustal x version 2.0. *bioinformatics*, 23(21):2947–2948. 35

Larsson, P., Wallner, B., Lindahl, E., and Elofsson, A. (2008). Using multiple templates to improve quality of homology models in automated homology modeling. *Protein Science*, 17(6):990–1002. 204

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography*, 26(2):283–291. 58

Le, Q., Sievers, F., and Higgins, D. G. (2017). Protein multiple sequence alignment benchmarking through secondary structure prediction. *Bioinformatics*, 33(9):1331–1337. 35

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., and Sheffler, W. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*, 487:545–574. 59

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–400. 162

Lee, D. A., Rentzsch, R., and Orengo, C. (2010). GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic acids research*, 38(3):720–737. 49, 66

Lee, G. R., Heo, L., and Seok, C. (2016). Effective protein model structure refinement by loop modeling and overall relaxation. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl 1):293–301. 54, 57, 88, 202

Lee, H., Baek, M., Lee, G. R., Park, S., and Seok, C. (2017). Template-based modeling and ab initio refinement of protein oligomer structures using galaxy in capri round 30. *Proteins: Structure, Function, and Bioinformatics*, 85(3):399–407. 147

Lees, J. G., Lee, D., Studer, R. A., Dawson, N. L., Sillitoe, I., Das, S., Yeats, C., Dessailly, B. H., Rentzsch, R., and Orengo, C. A. (2014). Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic acids research*, 42(D1):D240—-D245. 71

Lensink, M. F., Velankar, S., Kryshtafovych, A., Huang, S.-Y., Schneidman-Duhovny, D., Sali, A., Segura, J., Fernandez-Fuentes, N., Viswanath, S., Elber, R., et al. (2016). Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics*, 82(Suppl 1):323–348. 100, 101

Lensink, M. F., Velankar, S., and Wodak, S. J. (2017). Modeling protein–protein and protein–peptide complexes: Capri 6th edition. *Proteins: Structure, Function, and Bioinformatics*, 85(3):359–377. 100, 101, 146, 147, 205

Lensink, M. F. and Wodak, S. J. (2013). Docking, scoring, and affinity prediction in CAPRI. *Proteins: Structure, Function, and Bioinformatics*, 81(12):2082–2095. 100, 101

Levitt, M. (1983). Molecular dynamics of native protein: I. computer simulation of trajectories. *Journal of molecular biology*, 168(3):595–617. 55

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *Journal of molecular biology*, 226(2):507–533. 55

Levy, E. D., Pereira-Leal, J. B., Chothia, C., and Teichmann, S. A. (2006). 3D complex: a structural classification of protein complexes. *PLoS computational biology*, 2(11):e0020155. 93, 118

Levy, E. D. and Teichmann, S. (2013). Structural, evolutionary, and assembly principles of protein oligomerization. *Prog. Mol. Biol. Transl. Sci*, 117:25–51. 91

Lewis, T. E., Sillitoe, I., Andreeva, A., Blundell, T. L., Buchan, D. W. A., Chothia, C., Cozzetto, D., Dana, J. M., Filippis, I., Gough, J., et al. (2015). Genome3D: exploiting structure to help users understand their sequences. *Nucleic Acids Research*, 43(D1):D382–D386. 64

Li, J., Cao, R., and Cheng, J. (2015). A large-scale conformation sampling and evaluation server for protein tertiary structure prediction and its assessment in CASP11. *BMC bioinformatics*, 16(1):337. 38

Light, S. and Elofsson, A. (2013). The impact of splicing on protein domain architecture. *Current opinion in structural biology*, 23(3):451–458. 151

Liu, S. and Altman, R. B. (2003). Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic acids research*, 31(16):4828–4835. 151

Liu, T., Wang, Y., Eickholt, J., and Wang, Z. (2016). Benchmarking deep networks for predicting residue-specific quality of individual protein models in CASP11. *Sci Rep*, 6:srep19301. 59

Lobley, A., Sadowski, M. I., and Jones, D. T. (2009). pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, 25(14):1761–1767. 64

Ma, J., Wang, S., Wang, Z., and Xu, J. (2014). MRFalign: protein homology detection through alignment of Markov random fields. *PLoS computational biology*, 10(3):e1003500. 57, 88, 202

Madera, M. (2008). Profile comparer: a program for scoring and aligning profile hidden markov models. *Bioinformatics*, 24(22):2630–2631. 38

Maghrabi, A. H. A. and McGuffin, L. J. (2017). ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. *Nucleic Acids Research*, 45(W1):W416–W421. 60, 61, 63

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713. 204

Manavalan, B. and Lee, J. (2017). SVMQA: Support-vector-machine-based protein single-model quality assessment. *Bioinformatics (Oxford, England)*, 33(16):2496–2503. 60, 61

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, 6(12):e28766. 56, 202

Marsh, J. A. and Teichmann, S. A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. *Annual review of biochemistry*, 84:551–575. 94

Martin, A. C. R. (1996). Fitting was performed using the McLachlan algorithm (McLachlan, A.D., 1982 “Rapid Comparison of Protein Structres”, Acta Cryst A38, 871-873) as implemented in the program ProFit. 39

Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489. 165

Mashiach, E., Nussinov, R., and Wolfson, H. J. (2010). FiberDock: Flexible induced fit backbone refinement in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 78(6):1503–1519. 104, 147, 205

Mayr, J. A., Zimmermann, F. A., Horváth, R., Schneider, H.-C., Schoser, B., Holinski-Feder, E., Czermin, B., Freisinger, P., and Sperl, W. (2011). Deficiency of the mitochondrial phosphate carrier presenting as myopathy and cardiomyopathy in a family with three affected children. *Neuromuscular Disorders*, 21(11):803–808. 153

McGuffin, L. J. and Roche, D. B. (2010). Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*, 26(2):182–188. 69, 203

McLachlan, A. D. (1972). Repeating sequences and gene duplication in proteins. *Journal of molecular biology*, 64(2):417–437. 166

McLachlan, A. D. (1982). Rapid comparison of protein structures. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 38(6):871–873. 39

Melo, F. and Sali, A. (2007). Fold assessment for comparative protein structure modeling. *Protein Science*, 16(11):2412–2426. 82

Melo, F., Sánchez, R., and Sali, A. (2002). Statistical potentials for fold assessment. *Protein science*, 11(2):430–448. 62, 69

Méndez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003). Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins: Structure, Function, and Bioinformatics*, 52(1):51–67. 106, 121

Miller, S., Janin, J., Lesk, A. M., and Chothia, C. (1987). Interior and surface of monomeric proteins. *Journal of molecular biology*, 196(3):641–656. 162

Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R., and Weng, Z. (2007). Integrating statistical pair potentials into protein complex prediction. *Proteins: Structure, Function, and Bioinformatics*, 69(3):511–520. 113

Mishra, A., Rao, S., Mittal, A., and Jayaram, B. (2013). Capturing native/native like structures with a physico-chemical metric (pcSM) in protein folding. *Biochim Biophys Acta Proteins Proteomics*, 1834(8):1520–1531. 54

Moal, I. H., Moretti, R., Baker, D., and Fernández-Recio, J. (2013). Scoring functions for protein–protein interactions. *Current opinion in structural biology*, 23(6):862–867. 99

Modi, V., Xu, Q., Adhikari, S., and Dunbrack, R. L. (2016). Assessment of template-based modeling of protein structure in CASP11. *Proteins: Structure, Function and Bioinformatics*, 84(Suppl 1):200–220. 61

Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., and Kryshtafovych, A. (2014). Evaluation of residue-residue contact prediction in CASP10. *Proteins: Structure, Function and Bioinformatics*, 82(Suppl 2):138–153. 56

Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., and Kryshtafovych, A. (2015). New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins: Structure, Function, and Bioinformatics*, 82(Suppl 1):131–144. 56

Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2010). Protein–protein docking dealing with the unknown. *Journal of computational chemistry*, 31(2):317–342. 101

Mosca, R., Céol, A., and Aloy, P. (2012). Interactome3D: adding structural details to protein networks. *Nature methods*, 10(1):47–53. 93, 96, 98, 99, 102, 111, 112, 134

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2013a). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic acids research*, 42(D1):D374–D379. 97, 108, 113, 135

Mosca, R., Pons, T., Céol, A., Valencia, A., and Aloy, P. (2013b). Towards a detailed atlas of protein–protein interactions. *Current opinion in structural biology*, 23(6):929–940. 95

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2016). Critical assessment of methods of protein structure prediction (CASP) - progress and new directions in Round XI. *Proteins: Structure, Function and Bioinformatics*, 82(Suppl 1):4–14. 56

Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995). A large scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–iv. 55

Mount, D. W. (2008). Comparison of the pam and blosum amino acid substitution matrices. *Cold Spring Harbor Protocols*, 2008(6):pdb–ip59. 31

Mukherjee, S. and Zhang, Y. (2009). MM-align: a quick algorithm for aligning multiple chain protein complex structures using iterative dynamic programming. *Nucleic acids research*, 37(11):e83. 42, 121, 144

Mukherjee, S. and Zhang, Y. (2011). Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure*, 19(7):955–966. 102

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540. 50, 94

Nair, R., Liu, J., Soong, T.-T., Acton, T. B., Everett, J. K., Kouranov, A., Fiser, A., Godzik, A., Jaroszewski, L., Orengo, C., et al. (2009). Structural genomics is the largest contributor of novel structural leverage. *Journal of structural and functional genomics*, 10(2):181–191. 99

Nanda, A., Carson-Walter, E. B., Seaman, S., Barber, T. D., Stampfl, J., Singh, S., Vogelstein, B., Kinzler, K. W., and Croix, B. S. (2004). Tem8 interacts with the cleaved c5 domain of collagen alpha3 (vi). *Cancer Research*, 64(3):817–820. 154

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453. 32

Negroni, J., Mosca, R., and Aloy, P. (2014). Assessing the Applicability of Template-Based Protein Docking in the Twilight Zone. *Structure*, 22(9):1356–1362. 103, 105, 121, 144

Nooren, I. M. A. and Thornton, J. M. (2003). Diversity of protein–protein interactions. *The EMBO journal*, 22(14):3486–3492. 92

Norel, R., Petrey, D., Wolfson, H. J., and Nussinov, R. (1999). Examination of shape complementarity in docking of unbound proteins. *Proteins: Structure, Function, and Bioinformatics*, 36(3):307–317. 146

Nugent, T. and Jones, D. T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24):E1540–7. 56, 202

Ochoa-Montaño, B., Mohan, N., and Blundell, T. L. (2015). CHOPIN: a web resource for the structural and functional proteome of Mycobacterium tuberculosis. *Database*, 2015:bav026. 64

Ofran, Y. and Rost, B. (2003). Analysing six types of protein–protein interfaces. *Journal of molecular biology*, 325(2):377–387. 92

Olechnovič, K. and Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins: Structure, Function, and Bioinformatics*, 85(6):1131–1145. 59

Orengo, C., Jones, D. T., and Thornton, J. M. (2003). *Bioinformatics: genes, proteins and computers*. BIOS Scientific Oxford. 33, 40, 41

Orengo, C. A., Michie, A., Jones, S., Jones, D. T., Swindells, M., and Thornton, J. M. (1997). Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109. 47

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298. 38, 56, 202

Pais, F. S.-M., de Cássia Ruy, P., Oliveira, G., and Coimbra, R. S. (2014). Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*, 9(1):4. 35

Park, H., Bradley, P., Greisen Jr, P., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D., and DiMaio, F. (2016). Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation*, 20(12):6201–6212. 54, 57, 88, 202

Park, H., Lee, G. R., Heo, L., and Seok, C. (2014). Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments. *PloS one*, 9(11):e113811. 54

Petrey, D., Xiang, Z., Tang, C. L., Xie, L., Gimpelev, M., Mitros, T., Soto, C. S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., et al. (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins: Structure, Function, and Bioinformatics*, 53(S6):430–435. 55

Pfeiffenberger, E., Chaleil, R. A., Moal, I. H., and Bates, P. A. (2017). A machine learning approach for ranking clusters of docked protein-protein complexes by pairwise cluster comparison. *Proteins: Structure, Function, and Bioinformatics*, 85(3):528–543. 147

Pieper, U., Webb, B. M., Dong, G. Q., Schneidman-Duhovny, D., Fan, H., Kim, S. J., Khuri, N., Spill, Y. G., Weinkam, P., Hammel, M., et al. (2014). ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research*, 42(D1):D336–D346. 82

Pierce, B. and Weng, Z. (2007). ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins: Structure, Function, and Bioinformatics*, 67(4):1078–1086. 112, 113, 147

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B.-H., Vreven, T., and Weng, Z. (2014). ZDOCK server: interactive docking prediction of protein protein complexes and symmetric multimers. *Bioinformatics*, 30(12):1771–1773. 62

Pillmann, H., Hatje, K., Odronitz, F., Hammesfahr, B., and Kollmar, M. (2011). Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology. *BMC bioinformatics*, 12:270. 156

Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342. 191, 192, 195

Pohl, M., Bortfeldt, R. H., Grützmann, K., and Schuster, S. (2013). Alternative splicing of mutually exclusive exons—A review. *Biosystems*, 114(1):31–38. 148

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., and Clements, J. (2012). The Pfam protein families database. *Nucleic acids research*, 40(D1):D290–D301. 95

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, 7(1):95–99. 26, 58

Rangwala, H. and Karypis, G. (2010). Introduction to Protein Structure Prediction. *Introduction to Protein Structure Prediction*. 58

Rausell, A., Juan, D., Pazos, F., and Valencia, A. (2010). Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences*, 107(5):1995–2000. 49

Rawi, R., Whitmore, L., and Topf, M. (2010). CHOYCE: A web server for constrained homology modelling with cryoEM maps. *Bioinformatics*, 26(13):1673–1674. 89

Ray, A., Lindahl, E., and Wallner, B. (2012). Improved model quality assessment using ProQ2. *BMC bioinformatics*, 13(1):224. 59, 60, 69, 203

Read, R. J. and Chavali, G. (2007). Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):27–37. 45, 70

Redfern, O. C., Harrison, A., Dallman, T., Pearl, F. M. G., and Orengo, C. A. (2007). CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Computational Biology*, 3(11):e232. 44, 47

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 9(2):173–175. 38, 160

Rentzsch, R. and Orengo, C. A. (2013). Protein function prediction using domain families. *BMC Bioinformatics*, 14(Suppl 3):S5. 49

Rodrigues, J. P. and Bonvin, A. M. J. J. (2014). Integrative computational modeling of protein interactions. *FEBS Journal*, 281(8):1988–2003. 100

Sadowski, M. I. and Jones, D. T. (2007). Benchmarking template selection and model quality assessment for high resolution comparative modeling. *Proteins: Structure, Function, and Bioinformatics*, 69(3):476–485. 53

Sadreyev, R. and Grishin, N. (2003). Compass: a tool for comparison of multiple protein alignments with assessment of statistical significance. *Journal of molecular biology*, 326(1):317–336. 37, 49

Sali, A., Berman, H. M., Schwede, T., Trewhella, J., Kleywegt, G., Burley, S. K., Markley, J., Nakamura, H., Adams, P., Bonvin, A. M. J. J., et al. (2015). Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure (London, England : 1993)*, 23(7):1156–67. 90

Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Protein structure by distance analysis*, 64(3):779–815. 55, 102, 111, 134

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(Suppl 1):D449–D451. 105

Sánchez-Pla, A., Reverter, F., de Villa, M. C. R., and Comabella, M. (2012). Transcriptomics: mRNA and alternative splicing. *Journal of neuroimmunology*, 248(1):23–31. 150

Sarti, E., Zamuner, S., Cossio, P., Laio, A., Seno, F., and Trovato, A. (2013). BACHSCORE. A tool for evaluating efficiently and reliably the quality of large sets of protein structures. *Computer Physics Communications*, 184(12):2860–2865. 69

Schindler, C. E., Chauvot de Beauchêne, I., De Vries, S., and Zacharias, M. (2017). Protein-protein and peptide-protein docking and refinement using attract in capri. *Proteins: Structure, Function, and Bioinformatics*, 85(3):391–398. 146

Schmucker, D. and Chen, B. (2009). Dscam and dscam: complex genes in simple animals, complex animals yet simple genes. *Genes & development*, 23(2):147–156. 151

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., and Zipursky, S. L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6):671–684. 150

Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., and Friedberg, I. (2013). Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS computational biology*, 9(5):e1003063. 49

Schwede, T. (2013). Protein modeling: what happened to the “protein structure gap”? *Structure*, 21(9):1531–1540. 99

Shatsky, M., Nussinov, R., and Wolfson, H. J. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins: Structure, Function, and Bioinformatics*, 56(1):143–156. 104

Shen, M. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein science*, 15(11):2507–2524. 59, 62, 69, 112, 121, 203

Shi, Y., Fernandez-Martinez, J., Tjioe, E., Pellarin, R., Kim, S. J., Williams, R., Schneidman-Duhovny, D., Sali, A., Rout, M. P., and Chait, B. T. (2014). Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Molecular & cellular proteomics : MCP*, 13(11):2927–43. 90

Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein engineering*, 11(9):739–747. 43

Shoemaker, B. A., Zhang, D., Tyagi, M., Thangudu, R. R., Fong, J. H., Marchler-Bauer, A., Bryant, S. H., Madej, T., and Panchenko, A. R. (2012). IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic acids research*, 40(D1):D834–D840. 95, 98, 164

Siddiqui, A. S. and Barton, G. J. (1995). Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Science*, 4(5):872–884. 47

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539. 34, 35

Siew, N., Elofsson, A., Rychlewski, L., and Fischer, D. (2000). MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9):776–785. 45

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski,

R. A., Lee, D., Lees, J. G., et al. (2015). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research*, 43(D1):D376–D381. 67

Singh, A., Kaushik, R., Mishra, A., Shanker, A., and Jayaram, B. (2016). ProTSAV: A protein tertiary structure analysis and validation server. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1864(1):11–19. 59

Sinha, R., Kundrotas, P. J., and Vakser, I. A. (2012). Protein docking by the interface structure similarity: how much structure is needed? *PloS one*, 7(2):e31349. 117, 119

Skwark, M. J. and Elofsson, A. (2013). PconsD: ultra rapid, accurate model quality assessment for protein structure prediction. *Bioinformatics*, 29(14):1817–1818. 61

Skwark, M. J., Raimondi, D., Michel, M., and Elofsson, A. (2014). Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput Biol*, 10(11):e1003889. 56

Smith, T. F. and Waterman, M. S. (1981). Comparison of biosequences. *Advances in applied mathematics*, 2(4):482–489. 32

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics*, 21(7):951–960. 53, 111, 134

Söding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, 33(Suppl 2):W244–W248. 53

Soni, N. and Madhusudhan, M. (2017). Computational modeling of protein assemblies. *Current Opinion in Structural Biology*, 44:179–189. 147, 205

Soom, M., Gessner, G., Heuer, H., Hoshi, T., and Heinemann, S. H. (2008). A mutually exclusive alternative exon of slo 1 codes for a neuronal BK channel with altered function. *Channels*, 2(4):278–282. 156

Stein, A., Mosca, R., and Aloy, P. (2011). Three-dimensional modeling of protein interactions and complexes is going ‘omics. *Current opinion in structural biology*, 21(2):200–208. 99

Stephan, M., Möller, F., Wiehe, T., and Kleffe, J. (2007). Self-alignments to detect mutually exclusive exon usage. *In silico biology*, 7(6):613–621. 156

Sugnet, C. W., Kent, W. J., Ares, M., and Haussler, D. (2004). Transcriptome and genome conservation of alternative splicing events in humans and mice. In *Pacific Symposium on Biocomputing*, volume 9, pages 66–77. 149, 150

Swindells, M. B. (1995). A procedure for detecting structural domains in proteins. *Protein Science*, 4(1):103–112. 47

Szilagyi, A. and Zhang, Y. (2014). Template-based structure modeling of protein–protein interactions. *Current opinion in structural biology*, 24:10–23. 100

Tang, Z. Z., Sharma, S., Zheng, S., Chawla, G., Nikolic, J., and Black, D. L. (2011). Regulation of the mutually exclusive exons 8a and 8 in the CaV1. 2 calcium channel transcript by polypyrimidine tract-binding protein. *Journal of Biological Chemistry*, 286(12):10007–10016. 153

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigó, R., Briggs, S. P., and Bafna, V. (2007). Improving gene annotation using peptide mass spectrometry. *Genome research*, 17(2):231–239. 155

Taylor, W. R. (1986). The classification of amino acid conservation. *Journal of theoretical Biology*, 119(2):205–218. 28

Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *Journal of molecular biology*, 208(1):1–22. 39, 40, 41, 46, 47, 67, 110

Terwilliger, T. C. (2011). The success of structural genomics. *Journal of structural and functional genomics*, 12(2):43–44. 99

Thieman, J. R., Mishra, S. K., Ling, K., Doray, B., Anderson, R. A., and Traub, L. M. (2009). Clathrin regulates the association of PIPKIgamma661 with the AP-2 adaptor beta2 appendage. *J Biol Chem*, 284(20):13924–13939. 152

Thompson, J. D., Linard, B., Lecompte, O., and Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one*, 6(3):e18093. 35

Thorsen, K., Sørensen, K. D., Brems-Eskildsen, A. S., Modin, C., Gaustadnes, M., Hein, A.-M. K., Kruhøffer, M., Laurberg, S., Borre, M., Wang, K., et al. (2008). Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Molecular & Cellular Proteomics*, 7(7):1214–1224. 154

Torchala, M., Moal, I. H., Chaleil, R. A. G., Fernandez-Recio, J., and Bates, P. A. (2013). SwarmDock: a server for flexible protein–protein docking. *Bioinformatics*, 29(6):807–809. 101

Tovchigrechko, A. and Vakser, I. A. (2006). GRAMM-X public web server for protein–protein docking. *Nucleic acids research*, 34(Suppl 2):W310–W314. 101

Tress, M. (2013). Protein Tertiary Structures: Prediction from Amino Acid Sequences. In *eLS*, page 9780470015902.a0003040.pub2. John Wiley & Sons, Ltd, Chichester, UK. 54

Tress, M. L., Abascal, F., and Valencia, A. (2017). Alternative splicing may not be the key to proteome complexity. *Trends in biochemical sciences*, 42(2):98–110. 206

Tress, M. L., Bodenmiller, B., Aebersold, R., and Valencia, A. (2008). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome biology*, 9(11):1. 151, 155

Tress, M. L., Martelli, P. L., Frankish, A., Reeves, G. A., Wesselink, J. J., Yeats, C., Ísólfur Ólason, P., Albrecht, M., Hegyi, H., Giorgetti, A., et al. (2007). The implications of alternative splicing in the encode protein complement. *Proceedings of the National Academy of Sciences*, 104(13):5495–5500. 151

Tsai, C.-J., Lin, S. L., Wolfson, H. J., and Nussinov, R. (1996). A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *Journal of molecular biology*, 260(4):604–620. 93

Tuncbag, N., Keskin, O., Nussinov, R., and Gursoy, A. (2012). Fast and accurate modeling of protein protein interactions by combining template interface based docking with flexible refinement. *Proteins: Structure, Function, and Bioinformatics*, 80(4):1239–1249. 103

Tyagi, M., Hashimoto, K., Shoemaker, B. A., Wuchty, S., and Panchenko, A. R. (2012). Large-scale mapping of human protein interactome using structural complexes. *EMBO reports*, 13(3):266–271. 102

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419. 150

UniProt (2017). UniProt: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169. 24, 36, 48, 52

Uziela, K., Hurtado, D. M., Shu, N., Wallner, B., and Elofsson, A. (2017). ProQ3D: Improved model quality assessments using Deep Learning. *Bioinformatics*, 33(10):1578–1580. 59, 61, 202

Uziela, K., Wallner, B., and Elofsson, A. (2016). ProQ3: Improved model quality assessments using Rosetta energy terms. *Nature Publishing Group*, 6:srep33509. 59, 61, 202

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics*, 48(2):227–241. 35, 164

Vénien-Bryan, C., Li, Z., Vuillard, L., and Boutin, J. A. (2017). Cryo-electron microscopy and x-ray crystallography: complementary approaches to structural biology and drug discovery. *Acta Crystallographica Section F: Structural Biology Communications*, 73(4):174–183. 27

Verardi, R., Traaseth, N. J., Masterson, L. R., Vostrikov, V. V., and Veglia, G. (2012). Isotope labeling for solution and solid-state nmr spectroscopy of membrane proteins. In *Isotope labeling in Biomolecular NMR*, pages 35–62. Springer. 26

Vreven, T., Hwang, H., Pierce, B. G., and Weng, Z. (2012). Prediction of protein–protein binding free energies. *Protein Science*, 21(3):396–404. 143

Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., Chaleil, R., Jimenez-Garcia, B., Bates, P. A., Fernandez-Recio, J., et al. (2015). Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of molecular biology*, 427(19):3031–3041. 113

Wallner, B. and Elofsson, A. (2003). Can correct protein models be identified? *Protein science*, 12(5):1073–1086. 59

Wallner, B. and Elofsson, A. (2005). All are not equal: a benchmark of different homology modeling programs. *Protein Science*, 14(5):1315–1327. 55

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology*, 13(1):e1005324. 56

Wang, X., Li, G., Yang, Y., Wang, W., Zhang, W., Pan, H., Zhang, P., Yue, Y., Lin, H., Liu, B., et al. (2012). An RNA architectural locus control region involved in Dscam mutually exclusive splicing. *Nature communications*, 3:1255. 150

Warburg, O. (1956). On the origin of cancer cells. *Science*, 123(3191):309–314. 198

Ward, A. B., Sali, A., and Wilson, I. A. (2013). Integrative structural biology. *Science (New York, N.Y.)*, 339(6122):913–5. 90

Webb, E. C. (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Number Ed. 6. Academic Press. 50, 66

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta Jr., S., Weiner, P., Profeta, S., and Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784. 58, 59

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587. 37, 155

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009). SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research*, 37(Suppl 1):D380–D386. 51, 64

Winter, C., Henschel, A., Kim, W. K., and Schroeder, M. (2006). SCOPPI: a structural classification of protein–protein interfaces. *Nucleic acids research*, 34(Suppl 1):D310–D314. 108

Wong, N., Ojo, D., Yan, J., and Tang, D. (2015). PKM2 contributes to cancer metabolism. *Cancer letters*, 356(2):184–191. 198

Xie, X., Liu, X., Zhang, Q., and Yu, J. (2014). Overexpression of collagen vi alpha3 in gastric cancer. *Oncology letters*, 7(5):1537–1543. 154

Xu, D., Lin, S. L., and Nussinov, R. (1997). Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *Journal of molecular biology*, 265(1):68–84. 146

Xu, J. and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics*, 26(7):889–895. 45

Yamada, K. D., Nishi, H., Nakata, J., and Kinoshita, K. (2016). Structural characterization of single nucleotide variants at ligand binding sites and enzyme active sites of human proteins. *Biophysics and physicobiology*, 13:157–163. 166, 181

Yamakawa, K., Huo, Y.-K., Haendel, M. A., Hubert, R., Chen, X.-N., Lyons, G. E., and Korenberg, J. R. (1998). Dscam: a novel member of the immunoglobulin superfamily maps in a down syndrome region and is involved in the development of the nervous system. *Human molecular genetics*, 7(2):227–237. 151

Yan, R., Xu, D., Yang, J., Walker, S., and Zhang, Y. (2013). A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Scientific reports*, 3:srep02619. 53

Yan, Y., Wen, Z., Wang, X., and Huang, S.-Y. (2017). Addressing recent docking challenges: A hybrid strategy to integrate template-based and free protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 85(3):497–512. 147

Yang, J., Wang, Y., and Zhang, Y. (2016). ResQ: an approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *J Mol Biol*, 428(4):693–701. 57

Yang, J., Zhang, W., He, B., Walker, S. E., Zhang, H., Govindarajoo, B., Virtanen, J., Xue, Z., Shen, H.-B., and Zhang, Y. (2015). Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl 1):233–246. 63

Yeats, C., Redfern, O. C., and Orengo, C. (2010). A fast and automated solution for accurately resolving protein domain architectures. *Bioinformatics*, 26(6):745–751. 159

Yu, J., Andreani, J., Ochsenbein, F., and Guerois, R. (2017). Lessons from (co-) evolution in the docking of proteins and peptides for capri rounds 28–35. *Proteins: Structure, Function, and Bioinformatics*, 85(3):378–390. 146

Zanussi, S., Doliana, R., Segat, D., Bonaldo, P., and Colombatti, A. (1992). The human type VI collagen gene. mRNA and protein variants of the alpha 3 chain generated by alternative splicing of an additional 5-end exon. *Journal of Biological Chemistry*, 267(33):24082–24089. 154

Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic acids research*, 31(13):3370–3374. 45, 70

Zemla, A., Venclovas, Č., Moult, J., and Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins: Structure, Function, and Bioinformatics*, 37(Suppl 3):22–29. 45, 70

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., and Hunter, T. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):556–560. 98, 102

Zhang, Q. C., Petrey, D., Garzón, J. I., Deng, L., and Honig, B. (2013). PrePPI: a structure-informed database of protein–protein interactions. *Nucleic acids research*, 41(D1):D828–D833. 95, 98

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710. 41, 45

Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, 33(7):2302–2309. 39, 42, 43, 70, 114

Zhou, H. and Skolnick, J. (2011). GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical journal*, 101(8):2043–2052. 59, 69