Remote protein homology as a spyglass into the past

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Klaus Oliver Kopec

aus Neuenbürg, Deutschland

Tübingen
2014

Tag der mündlichen Qualifikation:	23.02.2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Andrei Lupas

ERKLÄRUNG

Hiermit erkläre ich, dass ich die Arbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind.

Tübingen, 2014

Klaus Oliver Kopec

ABSTRACT

Contemporary proteins evolved by the acquisition, recombination, and adaptation of established building blocks called domains. Most domains already existed at the time of the last universal common ancestor and then diverged, leaving weak signatures of their *homology*, i. e. common ancestry. Yet, the rapid growth of protein databases and improved algorithms revealed distant homologies of domains hitherto deemed unrelated. Besides evolutionary implications, these findings also enable the transfer of knowledge between similar proteins. It is important to probe the strengths, weaknesses, and limits of these methods to leverage this progress. To this end, we analyzed remote homologs in bioinformatic case studies.

First, we analyzed the uncharacterized SMP domain abundant in the ERMES complex, which tethers the endoplasmic reticulum to mitochondria and positively impacts inter-membrane phospholipid transfer. We established the BPI- and Takeout-like families as SMP domain homologs. As both families comprise hydrophobic ligand binders and share a fold, we predicted the same fold and an active role in phospholipid transfer for SMP domains. Finally, we grouped the three families in the novel tubular lipid-binding proteins (TULIP) superfamily.

Next, we searched for different folds with homology to the repetitive subunit of β-propellers, the *blade*, and initially detected four candidates. Further evaluation confirmed blade homologs in type II β-prism and IRE1-LD proteins, but revealed that WW domains and β-pinwheels have arisen convergently. These findings stress the importance of fold-spanning relationships for classification systems.

Lastly, we analyzed the tetratrico peptide repeat (TPR) motif to probe the lower sequence length limit in homology detection. We were unable to expand on known TPR homologs and our data did not allow us to infer an evolutionary scenario for TPR-like motifs. Due to a discussed origin from single motif instances, we also searched for TPR domains recently amplified from non-repetitive singleton instances but were unable to detect any and thus believe that this process is not ongoing.

Overall, our results help to determine more clearly the uses and limits of remote homology detection algorithms. Blade-sized fragments are within reach of current methods, whereas TPRs are already borderline cases. Further, the remote homologies uncovered in this work contribute to a growing knowledge base on protein evolution, which will eventually lead to a protein classification by natural descent.

Die heutigen Proteine entwickelten sich durch Neuerwerb, Rekombination und Anpassung etablierter Komponenten, den Domänen, weiter. Diese existierten bereits zur Zeit des letzten gemeinsamen Vorfahren (engl. last universal common ancestor, LUCA) und divergierten dann. Heute ist der gemeinsame Ursprung dieser Domänen - ihre Homologie - häufig kaum noch zu erkennen. Mit zunehmender Größe von Proteindatenbanken sowie verbesserter Algorithmen können jedoch heute Homologien zwischen Domänen erkannt werden, die vormals als unverwandt angesehen wurden. Neben ihrer evolutionären Bedeutung erlauben es diese Entdeckungen auch, Wissen zwischen ähnlichen Proteinen zu transferieren. Um die Möglichkeiten und Limitierungen dieser Methoden zu ermitteln, analysierten wir entfernte Homologe in bioinformatischen Fallstudien.

In einer ersten Studie analysierten wir die SMP Domäne, welche bis zu diesem Zeitpunkt nicht genauer beschrieben war. Sie ist Teil mehrerer Proteine des ERMES Komplexes, der das endoplasmatische Retikulum und die Mitochondrien verbindet und am Phospholipidtransfer zwischen den Membranen beteiligt ist. Wir konnten zeigen, dass diese Domäne homolog zu den Proteinen der BPI- und Takeout-ähnlichen Familien ist. Da diese beiden Familien sich in ihrer Funktion, dem Binden hydrophober Liganden, ähneln und dieselbe Faltung annehmen, sagten wir für SMP Domänen diese Faltung und eine aktive Rolle im Phospholipidtransfer vorher. Aufgrund ihrer Homologie gruppierten wir die drei Familien in der neudefinierten TULIP Superfamilie (engl. tubular lipid-binding proteins).

Danach suchten wir nach Homologen der repetitiven Untereinheit der β-Propeller Faltung, dem *Blatt*, und fanden vier solcher Faltungen. Weitergehende Untersuchungen bestätigten Homologe von Blättern in Typ II β-Prismen und IRE1-LD Proteinen, zeigten jedoch auch, dass es sich bei WW Domänen und β-pinwheel Proteinen um konvergente Entwicklungen handelt. Diese Erkenntnisse unterstreichen die Bedeutung von faltungsübergreifenden Homologien für Proteinklassifikationssysteme.

Zuletzt untersuchten wir TPR-Motive, um uns der Mindestlänge für Homologiesuchen anzunähern. Wir konnten keine unbekannten Homologe von TPR-Motiven finden und unsere Daten waren nicht ausreichend, um ein Szenario der evolutionären Abstammung der TPR-ähnlichen Motive zu entwickeln. Da eine mögliche Abstammung der TPR-Motive von einer einzelnen Motivinstanz diskutiert wurde, suchten wir auch nach kürzlich amplifizierten TPR Domänen, die nicht von TPR-Motiven aus einem repetitiven Kontext abstammen. Wir konnten

keinen solchen Fall identifizieren und schlossen daraus, dass dieser Prozess heute nicht mehr stattfindet.

Insgesamt helfen unsere Resultate bei der Bestimmung von Nutzen und Grenzen moderner Homologieerkennungsmethoden. Fragmente in der Größenordnung von β-Propellerblättern können mit diesen Methoden analysiert werden, wohingegen deren Potential für TPR-Motive eingeschränkt ist. Die von uns entdeckten Homologien tragen zudem zum wachsenden Wissen über Proteinevolution bei, das letztendlich den Weg zu einer abstammungsbasierten Proteinklassifikation ebnet.

What's past is prologue.

— William Shakespeare

ACKNOWLEDGMENTS

I would like to thank Prof. Dr. Andrei Lupas for his expertise and continuous support which made this work possible. His suggestions regarding project planning and realization as well as publication writing will be of great value to me beyond my work in his department. Further, I express my gratitude to Prof. Dr. Oliver Kohlbacher for the valuable feedback on my research projects and dissertation. I also thank my thesis advisory committee members Dr. Birte Höcker and Dr. Oliver Weichenrieder for their questions and recommendations regarding my research and its presentation.

In addition, I would like to especially thank Dr. Vikram Alva who sacrificed countless hours of his time introducing me to protein evolution bioinformatics and the beauty of scientific writing. Neither my research nor this dissertation are conceivable without him!

Dr. Martin Mechelke, Martin Schückel, and Jens Baßler deserve a mentioning for making my workdays much more enjoyable than they would have been otherwise.

Finally, my wife Monika has my eternal gratitude for supporting me throughout my doctorate.

CONTENTS

LIST OF FIGURES

Chapter C

LIST OF TABLES

ACRONYMS

al2co	a program for scoring the conservation of multiple alignment columns.
BP1	type I β-prism, an all-β protein fold with three-fold symmetry and β-strands running parallel to the pseudo-symmetry axis.
BP2	type II β-prism, an all-β protein fold with three-fold symmetry and β-strands running orthogonal to the pseudo-symmetry axis.
BPI	bactericidal/permeability-increasing proteins, a protein with two structurally very similar tandem domains with hydrophobic ligand-binding propensity in a central tunnel.
BLAST	basic local alignment search tool, a heuristic homology detection algorithm based on a seed-and-extend approach.
BLOSUM	block substitution matrix, an amino acid substitution matrix derived from ungapped blocks of aligned homologous proteins.
BLOSUM62	BLOSUM derived from blocks with at least 62% sequence identity.
CASP	critical assessment of protein structure prediction, a competition for assessing the protein modeling state-of-the-art.
CATH	class, architecture, topology, homology, a hierarchical protein classification system using structure and sequence data.
CETP	cholesteryl ester transfer protein, a protein structurally similar and related to BPI.
CLANS	cluster analysis of sequences, a software to produce a graph layout of protein sequences based on their pairwise similarities.
CSB	computational structural biology toolbox, a Python library for computational structural biology.
CstF	cleavage stimulating factor

CstF-77 cleavage stimulating factor 77, the 77kDa subunit of the heterotrimeric CstF complex.

DHDPS dihydrodipicolinate synthase, a family of proteins that adopt a TIM barrel fold.

DNA deoxyribonucleic acid, the macromolecule on which the genome of an organism is encoded and passed on to descendants.

ER endoplasmic reticulum, an organelle of eukaryotic cells.

ERMES endoplasmic reticulum-mitochondria encounter structure, a protein complex tethering ER to mitochondria in yeast.

ESAG5 expression site-associated gene 5, a group of proteins from *Giardia* that resembles BPI.

ESCRT-III endosomal sorting complexes required for transport III

E-SYT2 extended synaptotagmin 2, a protein comprising an SMP domain; tethers ER and plasma membrane.

FASTA a sequence comparison algorithm nowadays primarily known for the file format of the same name.

Get guided-entry of tail-anchored proteins, a pathway targeting proteins to the ER membrane.

GNA *Galanthus nivalis* agglutinin, a family of proteins that bind sugars.

HAT half-a-TPR, a variant of TPRs with different conservation pattern.

HEPN higher eukaryotes and prokaryotes nucleotide-binding, the C-terminal domain of chaperonin sacsin.

HHalign a program for pairwise profile HMM alignment. Part of HH-suite.

HHblits HMM-HMM-based lightning-fast iterative sequence search, a program for iterative pairwise profile HMM comparisons that uses fast pre-filtering for speed-up. Part of HH-suite.

HHmake a program for converting MSAs to HH-suite format profile HMMs. Part of HH-suite.

HHpred a webserver for pairwise profile HMM comparisons with the possibility of subsequent structure modeling. Internally uses, e. g., a modified PSI-BLAST, HHblits, HHmake, HHsearch, and HHalign.

HHrepID a program for repeat detection based on the comparison of a profile HMM to itself.

HHsearch a program for pairwise profile HMM comparisons. Part of HH-suite.

HHsenser a program for iterative homology detection with result verification by profile HMMs.

HH-suite a range of programs relevant for pairwise profile HMM comparisons. Includes, eg HHalign, HHblits, HHmake, and HHsearch.

HMM hidden Markov model, a probabilistic model widely used in bioinformatics.

HMMER a program for comparing profile HMMs to sequences.

HSP	high-scoring segment pair, a pair of regions in two sequences aligned by BLAST with a score above threshold.
IRE1-LD	inositol-requiring enzyme 1 luminal domain, a eukaryotic protein in the unfolded protein response pathway.
I-TASSER	a fold prediction metaserver. Top server in CASP 7 & 8.
JHBP	juvenile hormone-binding protein, an insect protein that binds hydrophobic ligands and especially a hormone.
LPLUNC	long PLUNC, tandem domain version of PLUNC.
LPSBP	lipopolysaccharide-binding protein
LSO	log-sum-of-odds, a score used in pairwise profile HMM comparisons.
LUCA	last universal common ancestor, a hypothetical organism preceding today's three kingdoms of life.
MEN1	multiple endocrine neoplasia type 1, a hereditary tumor syndrome affecting endocrine organs.
Mmm1	maintenance of mitochondrial morphology protein 1, a protein in the ERMES complex.
Mmm2	maintenance of mitochondrial morphology protein 2, another name for Mdm34.
Mdm10	mitochondrial distribution and morphology protein 10, a protein in the ERMES complex.
Mdm12	mitochondrial distribution and morphology protein 12, a protein in the ERMES complex.
Mdm34	mitochondrial distribution and morphology protein 34, a protein in the ERMES complex, sometimes named Mmm2.
MIM	MIT interacting motif, a motif in ESCRT-III subunits; recognized and bound by MIT domains.
MIT	microtubule interacting and trafficking, a three-helix bundle domain; binds MIMs.
MRF	Markov random field, a generalization of HMMs to dependencies between arbitrary states.
MSA	multiple sequence alignment, an assignment of homologous residues to each other for more than two sequences.
MULTICOM	a fold prediction server.
MUSTER	a fold prediction server.
NCBI	National Center for Biotechnology Information, one of the main resources for biomedical and genomic information.
nr	a non-redundant protein sequence database, i. e. only one of identical sequences is retained.
nr20	nr clustered at 20% sequence identity.
OMP	outer membrane protein

PDB protein data bank, the primary databank for experimentally solved protein structures.

PDB70 PDB clustered at 70% sequence identity.

Pfam a database classifying proteins into families and more coarsely into clans.

Phyre a fold prediction server.

PLTP phospholipid transfer protein

PLUNC palate, lung, and nasal epithelium carcinoma-associated protein, a family of proteins related to BPI and CETP that exists in short and long forms called SPLUNC and LPLUNC, respectively.

PP5 protein phosphatase 5

PPR pentatrico peptide repeat, a 35 residue $\alpha\alpha$-hairpin forming solenoid domains if repeated; related to TPR.

PQQ pyrroloquinoline quinone, a small molecule acting as cofactor in, e. g., some β-propellers.

PSI-BLAST position-specific iterated BLAST, an iterative BLAST version that uses PSSM for scoring.

PSSM position-specific scoring matrix, a substitution matrix built from position-specific amino acid frequencies of an MSA.

PTS phosphotransferase system, a sugar-based bacterial signal transduction network.

RMSD root-mean-square deviation, a score for pairwise protein structure similarity based on aligned residues.

RNA ribonucleic acid, a macromolecule used in, e. g., information carrier and gene expression regulator.

RP 19S regulatory particle, a part of the 26S proteasome.

Rpn RP non-ATPase

RPS20 ribosomal protein S20, a ribosomal protein comprising a single TPR-like $\alpha\alpha$-hairpin.

SCOP structural classification of proteins, a hierarchical protein classification system using structure and sequence data.

SCOP70 structural classification of proteins (SCOP) clustered at 70% sequence identity, often obtained from the ASTRAL compendium on the SCOP website.

SEL1-like a repeat motif; related to TPR.

SMART simple modular architecture research tool, a resource for annotating known domains types in a protein.

SMP synaptotagmin-like, mitochondrial, and lipid-binding proteins, a family of proteins found, e. g., in the ERMES complex.

SPLUNC short PLUNC, single domain version of PLUNC.

SSE secondary structural element, elements on the secondary protein structure level.

super-SSE	supersecondary structural element, a compact local arrangement of multiple SSE.
Sus	starch utilization system, a sugar-processing complex of *Bacteroidetes*.
SusD	starch utilization system protein D, a sugar-processing protein of *Bacteroidetes*.
THA8L	thylakoid assembly 8-like
TIM	triosephosphate isomerase, a $(\beta\alpha)_8$-barrel fold.
TOM20	translocase of outer membrane 20kDa subunit, a protein of the TOM complex that transfers proteins through mitochondrial membranes.
TPR	tetratrico peptide repeat, a 34 residue $\alpha\alpha$-hairpin forming solenoid domains if repeated; related to PPR and SEL1-like motifs.
TPRpred	a modified version of HHsearch fine-tuned for the peculiarities of TPR, PPR, and SEL1-like motif detection.
TM-align	a program for computing a structural alignment that is (close to) optimal with respect to the TM-score.
TM-score	a score for pairwise protein structure similarity based on aligned residues that is independent of alignment length. Optimized by TM-align.
TRUST	a repeat detection algorithm.
TULIP	tubular lipid-binding proteins, a superfamily of proteins presumably binding hydrophobic ligands.

1

INTRODUCTION

Most of the independently-folding subunits of contemporary proteins, the domains, probably already existed at the time of the last universal common ancestor (LUCA). In the post-LUCA era, new proteins were mostly formed by acquisition, recombination, and adaptation of these building blocks, making them the prime unit of recent evolution (Chothia *et al.*, 2003). These domains gradually diversified and thereby spawned families of related yet distinct proteins. The proteins within one family hence descend from a common ancestor; they are *homologs*. Due to neutral drift and other mutations, homologs can be quite different, often to a point where their sequence similarity is not significant any more (Kimura, 1968). It is nonetheless often possible to establish their relationship using advanced algorithms, enabling the reconstruction of ancestral features and evolutionary intermediates. Given that proteins do not fossilize and that ancestral organisms are often extinct, such inferences allow us to glance back into an otherwise inaccessible protein age.

The first step in these analyzes is to uncover homologs of the protein of interest. The comparison of sequences is the most reliable approach for establishing homology. The rationale behind this is that there is a large number of possible sequences already for moderate protein lengths, e. g., 20^{300} (~10^{390}) sequences are possible for the 20 proteinogenic amino acids and the average protein length of 300 residues. It is thus highly unlikely that two unrelated proteins converge to the same sequence, making sequence similarity the hallmark of homology. In contrast, owing to biophysical constraints, similar protein structures can arise by convergence and are therefore generally considered as analogous in absence significant sequence similarity (Krishna and Grishin, 2004).

Recently, the availability of high-throughput methods for sequencing has significantly increased the growth rate of protein sequence databases, necessitating fast approaches for comprehensive searches. The detection of the aforementioned homologs with little sequence similarity within these large datasets further depends on sensitive methods. In light of these two requirements, remote homology detection methods have evolved quickly in the last decades.

Initially, query and template sequences were compared directly and in many approaches, e. g., BLAST, fast heuristics sped up the process (Altschul *et al.*, 1990). BLAST is the standard application for basic homology searches, but it often fails to detect nontrivial relationships, limiting its use for more distant homologies. To alleviate

this problem, sequence profiles were introduced that harnessed the position-specific amino acid distribution encoded in an alignment of sequences related to the query. Another advantage of profiles is that they can be iteratively refined easily with newly detected homologs, as implemented in PSI-BLAST (Altschul *et al.*, 1997). Profile-sequence and later also the slower pairwise profile comparisons were shown to improve on the sensitivity of pairwise sequence comparisons (Sadreyev and Grishin, 2003). However, gaps are scored uniformly in sequence profile-based methods, which does not reflect the clustering of gaps often observed in certain alignment regions, e. g., those covering unconserved loops in the structure.

The position-specific scoring of insertions and deletions was introduced with profile HMMs, which are probabilistic models with parameters directly derivable from sequence alignments (Durbin *et al.*, 1998). Programs using profile HMMs are fast and sensitive and are therefore used in many well-known resources like the protein family database Pfam, which uses HMMER to compare query sequences to family profile HMMs (Finn *et al.*, 2014; Eddy, 2011). Analogous to the developments in sequence profile methods, pairwise profile HMM comparisons were introduced as part of HHpred and later HHblits (Söding *et al.*, 2005; Remmert *et al.*, 2012). The high sensitivity offered by these two programs is the state-of-the-art in remote homology detection, as evident from the top-scoring results of HHpred in assessments of template-based structure predictions for which homologous templates are key (Hildebrand *et al.*, 2009; Mariani *et al.*, 2011; Huang *et al.*, 2014).

These advances of remote homology detection methods resulted in a large number of proteins now shown to be homologous but previously deemed analogous. In a comprehensive study, Alva *et al.* (2010) performed pairwise profile HMM comparisons of the entries in the structural classification of proteins (SCOP) (Murzin *et al.*, 1995). They found that the different superfamilies of many folds are homologous. This is particularly interesting as the lack of evidence for their common ancestry led to their division into different superfamilies, i. e. they were initially considered convergent developments. Further, connections were found between superfamilies in seemingly unrelated folds. While different folds are generally considered as analogous developments, several fold-spanning relationships have been established based on homologous fold change (e. g., Grishin, 2001a; Andreeva and Murzin, 2006; Alva *et al.*, 2008) and conserved supersecondary structural elements (super-SSEs) (e. g., Copley *et al.*, 2001; Alva *et al.*, 2007; Coles *et al.*, 2006). To incorporate these findings into classification systems, the *metafolds* was recently proposed as a new classification level to group homologous folds (Alva *et al.*, 2008). Their automated detection in a high-throughput survey signifies an important step in remote homology detection by which analogous criteria

could eventually be purged from classification systems, resulting in a classification by natural descent.

Besides their usefulness in evolutionary studies, homologous information also makes the transfer of knowledge between proteins more reliable. This resembles the use of model organisms, from which knowledge can be transferred to other species due to the descent of life from a common ancestor. In protein research, comparable benefits are structure and function prediction, as they are often conserved even between remote homologs. This possibility is exploited in structural genomics, where diverse protein structures are solved to act as template for structure predictions (Burley *et al.*, 1999). Finally, experiments on proteins of unestablished organisms might be guided towards systems that are easier to study if homologs can be found.

To benefit from detected homologies in any of the aforementioned ways, it is important to be able to judge on the reliability of these findings in practical research applications. For the present work, we were therefore interested in probing the limits of current remote homology detection methods with case studies on different structural levels of proteins. We wanted to study the depth in which an uncharacterized domain could be analyzed and annotated with fold and function prediction without close well-studied relatives. The results of such a detailed account might also directly implicate an evolutionary trajectory leading to the domains we encounter today. To find a rough estimate of the sequence length at which homologies can still be reliably established, we would like to then move to smaller targets, the size of supersecondary structural elements. Overall, these analyzes would allow us to judge more precisely on the possibilities and limits of remote homology detection for small proteins or fragments. Further, our studies would contribute to the knowledge base on remote homologs and their properties. Ultimately, we hope to polish the spyglass lenses, allowing for a clearer view into the past.

The upcoming chapters cover the necessary background in protein evolution bioinformatics, three case studies of protein evolution, and their conclusions. The first analysis is concerned with the uncharacterized synaptotagmin-like, mitochondrial, and lipid-binding proteins (SMP) domain family, about which little, besides the sequences of a few members, was known when we started this project. Recently, two SMP domains were found in the endoplasmic reticulum-mitochondria encounter structure (ERMES) complex, which tethers endoplasmic reticulum (ER) and mitochondrial membranes together and is required for efficient inter-organelle phospholipid transfer. We tried to elucidate the origin, distribution, and evolution of SMP domains, as well as their function in general but also specifically in the context of ERMES. To this end, we charted the protein sequence space of SMP domain-like proteins using homology searches as well as clustering. We established a homologous relationship between SMP domains

and the bactericidal/permeability-increasing proteins (BPI)-like and Takeout-like protein families, which share a common fold and have lipid/hydrophobic ligand-binding properties. Based on thorough analyzes, we proposed the same fold and function for SMP domains and grouped all three families into the novel tubular lipid-binding proteins (TULIP) domain superfamily. Additionally, it became evident that members of the BPI-like family are the closest to the ancestral TULIP domain, which we substantiated with a data-derived evolutionary trajectory. This further led to a reasonable explanation for the predominance of two tandem TULIP domains in BPI-like proteins, whereas SMP domain-like and Takeout-like proteins mostly comprise only one. The implications of our detailed account of this superfamily have been widely recognized by the scientific community and the recently published structure of a lipid-bound an SMP domain confirmed our fold and function predictions.

In chapter four, we analyze the evolution of β-propeller proteins and their homologs of different folds. Proteins with β-propeller fold are thought to have arisen by repetition of a single ancestral *blade*, a four-stranded anti-parallel β-sheet characteristic for β-propellers. As, additionally, the high prevalence of β-propellers is indicative of their suitability as stable scaffold, we wanted to find out whether homologs with different folds arose from single blades or fully-formed β-propellers. Using homology searches and clustering, we then compiled a set of four different folds with significant sequence similarity to β-propeller blades. Thorough case-by-case analyzes incorporating the structural evolution in addition to sequence information eliminated two analogous candidates and revealed inositol-requiring enzyme 1 luminal domain (IRE1-LD) and proteins with the type II β-prism (BP2) fold as β-propeller homologs. Additional considerations led us to propose that IRE1-LD was derived from a complete 8-bladed β-propeller, whereas the BP2s fold probably arose independently by amplification from a single blade. Both are interesting examples of fold-spanning homologies. Motivated by our findings of the first non-β-propeller folds that are based on blades, we are currently investigating the design of a protein with BP2 fold from a β-propeller blade, which we intend to confirm experimentally.

In the final analysis, we explore the evolution of tetratrico peptide repeats (TPRs), an $\alpha\alpha$-hairpin motif that forms solenoid domains by repetition. Due to its proposed origin by amplification from a single motif instance, we tried to find recently amplified TPR domains, but could not detect any. This implies that novel TPR proteins are merely formed by adaptation of existing domains. Further, its short length of 34 residues makes this motif especially interesting for probing the length limits of contemporary remote homology detection. The compilation of a comprehensive dataset of TPR-like $\alpha\alpha$-hairpins was clearly hindered by the short motif length. Even after an elaborate

dataset assembly approach, the final dataset did not help us to analyze the evolution of TPRs in detail or to derive a trajectory of events for $\alpha\alpha$-hairpin motif evolution beyond what was already known. Our analysis of the dataset revealed that expert-annotated homologs were not deemed significantly similar and that analogs could not be distinguished with certainty from homologs. We therefore concluded that diverse sequences as short as TPRs are beyond the capabilities of pairwise hidden Markov model (HMM) comparisons with respect to the reliable detection of remote homologs and that novel methods might be required for this task.

2

BACKGROUND

This chapter includes content from the following publication.
Re-use permissions have been granted by the copyright holder.

Kopec K.O. and Lupas A.N.
β-Propeller Blades as Ancestral Peptides in Protein Evolution.
PLoS ONE, *8*(10): e77074, 2013.

2.1 PROTEINS

The diversity of proteins arises from the use of different combinations of 20 naturally occurring amino acids. These amino acids share a basic structure with an amino ($-NH_2$) and a carboxyl ($-COOH$) group connected to the same carbon atom, called C_α; they are thus α-amino acids. While the third covalent bond is to a hydrogen atom, the fourth substitutent is the *side-chain*, which causes different chemical and physical properties. All amino acids therefore have chiral C_α atoms, with the exception of glycine due to its second C_α hydrogen. Further, the proteinogenic amino acids are L-enantionmers (see figure 2.1 for an example), however a few notable cases of D-forms are known (e. g., Lam *et al.*, 2009; Wolosker *et al.*, 2008; Pisarewicz *et al.*, 2005).

Amino acids are covalently linked by peptide bonds during the biosynthesis of proteins. The result is a linear polypeptide chain with a repeating series of C_α-C-N atoms known as the *main chain* or *backbone* (figure 2.2). This linearity results in a free amino and carboxyl group at the two termini, which are named the N- and C-terminus, respectively. The number of subsequent amino acids, called *residues* in the context of the chain, varies greatly between proteins and can be as low as a few dozen and as high as several tens of thousands (Sudol *et al.*, 1995; Opitz *et al.*, 2003).

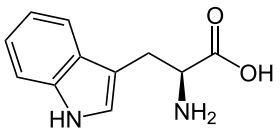


Figure 2.1: The amino acid L-tryptophan is an example of an α-amino acid with chiral α-carbon.

2.1.1 *Structural levels*

Proteins are usually described on four levels known as primary, secondary, tertiary, and quaternary structure (figure 2.3; Linderstrøm-Lang, 1952). The complexity encountered between the secondary and tertiary structural levels can be further detailed by supersecondary structural elements and domains.

The *primary structure* of a protein is the *sequence* of amino acid residues in its polypeptide chain. Sequences are commonly represented in 1-letter code, e. g., *G* represents glycine, and are reported from N- to C-terminus (Branden and Tooze, 1999). In bioinformatics, this representation is ubiquitous in the FASTA file format (figure 2.4).

The *secondary structure* describes chain segments with a regular local structure stabilized by mainchain-mainchain hydrogen bonds. The most encountered secondary structural element (SSE) are α-helices, β-sheets, and turns.

The α-helical geometry was correctly predicted several years before the first protein structure was solved (Pauling *et al.*, 1951; Kendrew *et al.*, 1958). In α-helices, mainchain-mainchain hydrogen bonds between CO and NH groups spaced by four residues ($i + 4 \rightarrow i$ hydrogen bonds) bring the backbone into a spring-like shape with ~3.6 residues per turn (Branden and Tooze, 1999). Amino acids differ in their propensity to be in a helical context, e. g., proline is known to break helices due to sterical hindrance and a lack of a hydrogen to donate for bonding (Pace and Scholtz, 1998).

A β-sheet is a parallel or antiparallel (or mix thereof) arrangement of neighboring extended chain segments called β-strands. Hydrogen bonds stabilize the alignment of neighboring β-strands and it is important to note that, unlike for α-helices, β-sheets do not necessarily consist of one continuous segment of the polypeptide chain but can comprise strands that are far apart in sequence. Similar to the situation in α-helices, residues have different β-strand propensities, which differ for initial and terminal residues of a strand and for internal and edge strands (Minor and Kim, 1994; Farzadfard *et al.*, 2008; Richardson and Richardson, 2002).

Finally, turns are two to six consecutive residues that alter the direction of the backbone, often stabilized by intra-turn hydrogen bond formation (Koch and Klebe, 2009). Different turn types are identified based on the number of involved residues and the hydrogen bond

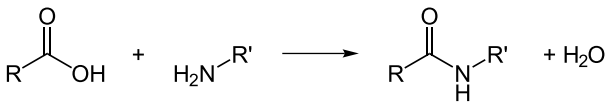


Figure 2.2: The condensation reaction between the carboxyl group of one and the amino group of another amino acid results in a peptide bond as well as the release of one molecule of water.

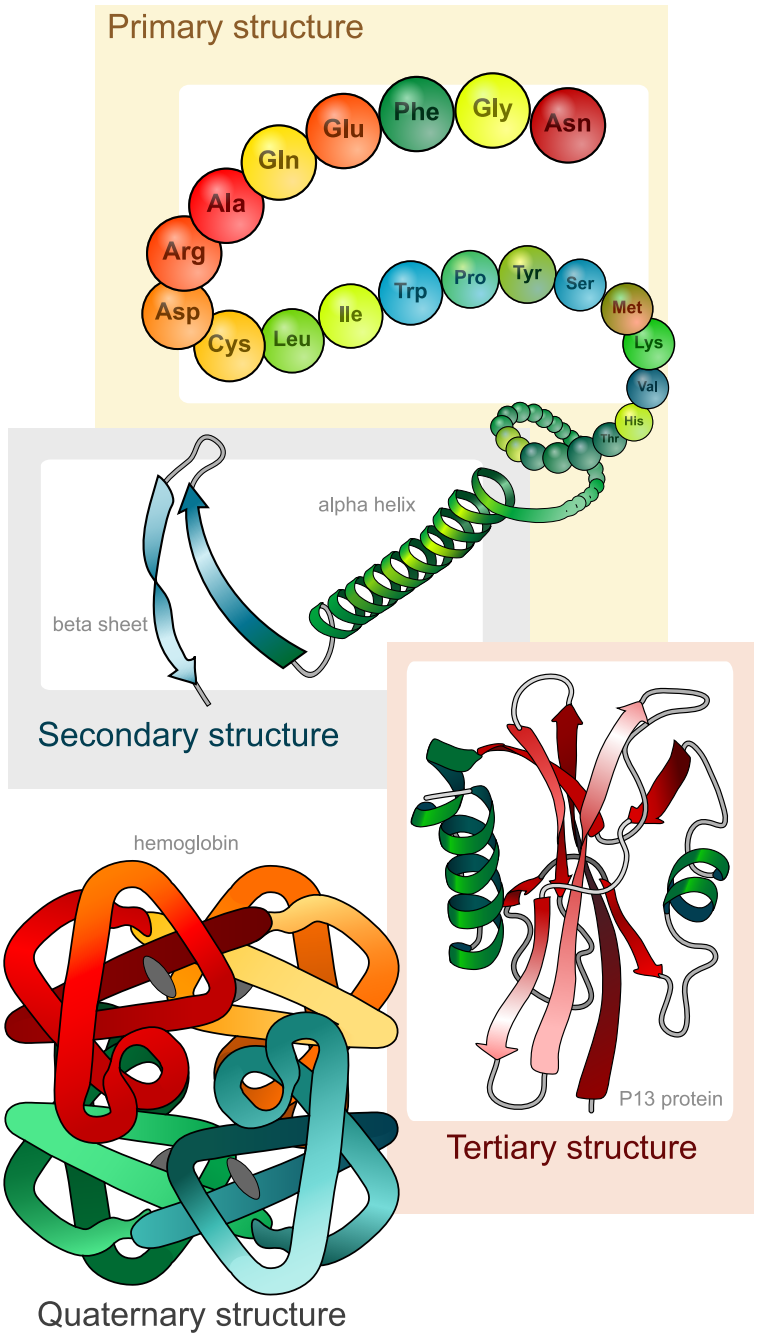


Figure 2.3: Symbolic depiction of the four main levels of protein structure. The primary structure is represented by a chain of balls, each labeled with an amino acid 3-letter code. The secondary structural elements and the tertiary structure are shown as cartoons. The abstractly visualized quaternary structure of the hemoglobin tetramer has colored domains and gray ovals indicating oxygen-binding heme groups. This figure is based on a public domain image by *LadyofHats* (`https://commons.wikimedia.org/wiki/User:LadyofHats`).

type, e. g., β-turns comprise four residues with a hydrogen bond between CO_i and NH_{i+3}. In general, turns can be seen as those SSEs that fold the chain back onto itself, leading to globular tertiary structures (Venkatachalam, 1968; Chou, 2000).

A supersecondary structural element (super-SSE) is a compact arrangements of multiple, spatially close SSEs. Most commonly found are $\alpha\alpha$-, $\beta\beta$-, and $\beta\alpha\beta$-hairpins (Salem *et al.*, 1999). The super-SSEs play an important role in the description of *domains* and the comparison of their folds.

Domains are the smallest independently folding subunits of a protein. The arrangement and connectivity of SSEs in a domain is called its *fold*. While folds can be very distinct, on average more than 60% are covered by the aforementioned, most common super-SSEs (Salem *et al.*, 1999). In the extremest of cases, repetitions of only one super-SSE constitute a fold, e. g., tetratrico peptide repeat (TPR) $\alpha\alpha$-hairpins form open-ended superhelical domains (Söding and Lupas, 2003). Folds can also be regarded as bipartite with defining *core* features and more variable embellishments; the former are required for fold membership while the latter can differ greatly between members. However, the assignment of residues to either subset can be difficult and classification systems thus often disagree (Csaba *et al.*, 2009).

The *tertiary structure* is the three-dimensional structure of a whole protein (Branden and Tooze, 1999). While identical to the domain for proteins with one domain, the tertiary structure of multi-domain proteins further comprises the domain arrangement and linker regions. The boundaries of domains can often be derived from the tertiary structure, however automatic domain assignments are difficult (Holland *et al.*, 2006).

The *quaternary structure* is the arrangement of multiple proteins in a complex (Branden and Tooze, 1999). A famous example is hemoglobin, which is a heterotetrameric complex with two different subunits that binds oxygen in blood (figure 2.3; Paoli *et al.*, 1996).

Most proteins need to *fold* to reached their functional form after they are translated by the ribosome, which means that they assume a stable conformation with low free energy (Bryngelson *et al.*, 1995). However, the number of theoretically possible conformations is al-

```
>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens GN=INS PE=1 SV=1
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

Figure 2.4: FASTA format for protein sequences. One FASTA entry comprises the header, a single line starting with '>' and with no fixed format or data requirements, and the sequence data in one or multiple lines with residues from N- to C-terminus in 1-letter code. Multiple entries can be stored in one file by concatenation.

ready very large for small proteins and even at very fast sampling rates it would take prohibitively long to sequentially sample all conformations (Levinthal, 1969). Instead, assisting chaperones as well as intermediate states ease the transition from the unfolded to the completely folded native state (Mashaghi *et al.*, 2014). The intermediates lower the free energy stepwise, e. g., by the burial of hydrophobic side-chains in the core and the formation of hydrogen bonds, thus bringing the conformation closer to the native state, which is a (local) minimum of the folding energy landscape (Bryngelson *et al.*, 1995). The correct folding of proteins is an essential process as evident from the diseases caused by the accumulation of misfolded proteins, e. g., Alzheimer's and Parkinson's disease (Selkoe, 2004).

2.1.2 *Evolution*

There are currently about 10^3 known folds and it seems unlikely that this number will increase dramatically (Andreeva *et al.*, 2008; Sillitoe *et al.*, 2013). Estimates are that 80% of domains assume one of only ~400 folds, which seems peculiar given the tremendous amount of possible—and actually encountered—sequences (Coulson and Moult, 2002). However, it becomes less surprising when considering the evolution of proteins and their domains.

The majority of contemporary proteins comprises multiple domains (Apic *et al.*, 2001). Most of these domains probably already existed at the time of the last universal common ancestor (LUCA), a hypothetical organism preceding the three domains of life (Doolittle, 1999). Protein domains are considered the main unit of recent, i. e. post-LUCA, evolution and their rearrangement and adaptation to novel contexts are thought to be the prime mechanisms of protein evolution (Björklund *et al.*, 2005; Pasek *et al.*, 2006; Moore *et al.*, 2008). But when domains initially evolved, other criteria were relevant for the evolutionary fitness, e. g., efficient folding, high stability, and capability to scaffold novel functions. A direct implication of the different fitness of folds is a non-uniform distribution of the domain population over available folds observed as the aforementioned imbalance.

For the evolution of domains themselves, it is attractive to consider an origin from a set of smaller building blocks or ancestral fragments, similar to how multi-domain proteins arose from single domains. These elements might have needed the context of other molecules to fold and the *ribonucleic acid (RNA) world* hypothesis assumes that RNA could have had the required properties (Jeffares *et al.*, 1998; Joyce, 2002). Possibly, the reach of exclusively RNA-based catalysis was extended to new reaction types by the binding of small abiotically-synthesized peptides (Fetrow and Godzik, 1998; Lupas *et al.*, 2001; Söding and Lupas, 2003; Orgel, 2004). The increased specificity achieved by structured peptides would have favored their selec-

tion, leading to an increase in peptides with high SSE-forming propensity. Early organisms likely had error-prone replication and peptide synthesis, leading to a great variability in sequence and length of the peptides. From this pool of peptides, some might have been able to fold independently and became the first proteins (Söding and Lupas, 2003).

The super-SSEs contained in these proteins were optimized for folding or maybe the catalysis of certain reactions. Due to their importance in early domains, remnants of these super-SSEs might still be detectable as ancestral fragments in contemporary proteins derived from these domains (Söding and Lupas, 2003). To this end, it is important to distinguish between similarity by divergence and convergence.

2.1.3 *Homology and analogy*

At the median protein length of 300 residues, there are $20^{300} (\sim 10^{390})$ different amino acid combinations, which makes it highly unlikely for unrelated proteins to converge to significant sequence similarity. Sequence similarity is therefore considered the hallmark of *homology*, i. e. the divergent origin from a common ancestor. In contrast, unrelated sequences can converge to the same fold owing to biophysical constraints and structural similarity thus is often considered an *analogous* trait (Murzin *et al.*, 1995; Sillitoe *et al.*, 2013). This distinction between analogy and homology allows for classifications that make the wealth of protein data more accessible.

The well known classification systems structural classification of proteins (SCOP) and class, architecture, topology, homology (CATH) organize proteins in a hierarchy with analogous criteria on the upper and homologous criteria on the lower levels (Murzin *et al.*, 1995; Sillitoe *et al.*, 2013). The first two levels of SCOP, *class* and *fold*, divide proteins by their SSE content and by different folds in one class. On the lower two levels, remote homologs form *superfamilies* and closely related subgroups within them become *families*. While SCOP is mainly based on manual assignments, CATH uses a semi-automatic approach to derive a comparable scheme. A consensus classification was recently proposed to average out errors introduced in either of these two classifications (Csaba *et al.*, 2009).

These classifications are tremendously useful, however their hierarchical structure limits their representation of homologous relationships to proteins that share a fold. And while even remote homologs often still adopt the same fold, instances of homologous fold change (Grishin, 2001a; Andreeva and Murzin, 2006; Andreeva *et al.*, 2007; Alva *et al.*, 2008) and homology based on conserved super-SSEs in different folds are known (Alva *et al.*, 2008, 2007; Grishin, 2001b; Copley *et al.*, 2001; Coles *et al.*, 2006). To alleviate this issue, the *metafold*

was proposed as novel classification level for fold-spanning homology (Alva *et al.*, 2008). If a grouping of proteins including their metafold existed, all analogous criteria could be removed from the classification to yield a classification by natural descent. To reach this goal, it is important to investigate and understand remote homology and probe the limits of available homology detection methods.

2.2 HOMOLOGY DETECTION

The most reliable way of detecting homologous relationships between proteins is to compare their sequences. As mentioned before, proteins with similar sequences are likely to have originated from a common ancestor, while the same cannot be stated for those similar in structure (section 2.1.3). The search for homologs of a query sequence thus depends on the computation of similarities between the query and a set of database proteins. These similarity scores are usually derived from a pairwise alignment of sequences.

2.2.1 *Alignments and similarity scores*

Sequence alignments are assignments of residues corresponding to each other in two or more sequences—for the sake of simplicity the following explanation focuses on the alignment of two sequences. Alignments can be represented by writing the 1-letter code sequences above each other with corresponding residues in the same column (figure 2.5). Residues that cannot be assigned are aligned to a newly introduced *gap* character in the other sequence, often a dash. While it is relatively easy to align closely related or otherwise similar sequences, extensive gapping can be required with very different sequences and the result may be questionable with respect to biological meaning. Given an alignment, the similarity between its sequences can be computed.

A common approach for calculating the similarity between two sequences given an alignment is to accumulate the individual scores of

```
                 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16    17 18 19 20 21 22 23 24 25 26 27 28 29
sequence_A       GIVEQCCTSICSLYQL---CNVIALKITYRAE
sequence_B       GIVEQCCASVCSLYQLENYCNVIA--ITYRAE
consensus        ******* * ******   *****  ******
```

Figure 2.5: An exemplary alignment of two artificial sequences *A* and *B* in 1-letter code. Positions with respect to sequence *A* are on top. Identical residues are highlighted as white letters on black background and indicated as * in the consensus line. Dashes are used as gap symbols.

all columns. The similarity scores necessary for these pairwise amino acid comparisons are looked up in pre-computed *substitution matrices*. The widespread block substitution matrices (BLOSUM) are derived from gapless alignments of homologous proteins and capture the amino acid variation within these alignments (Henikoff and Henikoff, 1992). BLOSUM matrices exist for different levels of conservation and represent the amino acid changes to be expected at the corresponding evolutionary distance, e. g., BLOSUM62 is based on alignments of sequences with at least 62% identity (Henikoff and Henikoff, 1992). Residues aligned to gaps are penalized separately as substitution matrices only consider amino acid exchanges. Assuming reasonable substitution matrices, the remaining challenge in homology detection is finding a meaningful alignment between the query and all database sequences within feasible time.

2.2.2 *Pairwise sequence comparison*

The initial approaches guaranteed optimality with respect to a given substitution matrix and gap score. The famous global alignment algorithm by Needleman and Wunsch (1970) and its derivate for local alignments by Smith and Waterman (1981) make use of dynamic programming (Eddy, 2004). Efficient implementations of these methods have quadratic runtime, which is prohibitively slow for searching large databases, as each database entry must be aligned to the query.

The next step in homology detection thus was to sacrifice optimality in favor of much faster heuristic methods like FASTA (Pearson and Lipman, 1988) and the basic local alignment search tool (BLAST) (Altschul *et al.*, 1990). While FASTA is nowadays mostly known for the file format of the same name (figure 2.4), BLAST has become the de facto standard for basic homology searches. BLAST applies a *seed-and-extend* approach and initially searches for short perfectly-matching regions, the *seeds*. The seeds are then extended to both sides until dissimilar residues are encountered, at which point the local alignment is considered complete and named a high-scoring segment pair (HSP). An E-value (expectation value) is computed for each HSP to report on the expected number of random sequences with at least the score of this HSP in a database of the given size. Reliable matches therefore are recognizable by E-values close to 0. The direct sequence comparisons used in the Needleman-Wunsch, Smith-Waterman, and BLAST algorithms work well for closely related and hence reasonably similar sequences. But they are not sensitive enough to detect more remote homologs.

2.2.3 *Profile-sequence comparison*

Sensitivity was further improved by incorporating the information of multiple sequences instead of just one (Altschul *et al.*, 1997). Sequence *profiles*, or position-specific scoring matrices (PSSM), capture the variability of several sequences through the position-specific amino acid frequencies of their multiple sequence alignment (MSA), which are stored as $20 \times N$ matrices comprising one row for each amino acid and one column for each of N alignment positions. The initial profile for position-specific iterated BLAST (PSI-BLAST) is calculated from the alignment of significant matches of one round of BLAST (Altschul *et al.*, 1997). Subsequent iterations of the algorithm use this profile instead of a substitution matrix to score amino acid exchanges differently for each query position. The values in the profile do not remain static but are updated after each iteration with newly detected, significant matches. It is important to note that gaps are still scored identically independent of their position in the sequence. Newer algorithms perform pairwise profile comparisons for enhanced sensitivity, however the additional costs in terms of runtime are high (Yona and Levitt, 2002; Sadreyev and Grishin, 2003; Söding, 2005).

2.2.4 *Profile HMM-sequence comparison*

State-of-the-art homology detection algorithms use profile HMMs instead of sequence profiles to include position-specific gap information. To arrive at a description of how profile HMMs are compared to sequences, we first introduce the basics of general HMMs. Next, profile HMMs are discussed (see page 18). And finally, we will return to the actual topic of this section, the comparison of hidden Markov models (HMMs) to sequences (see page 20).[1]

HIDDEN MARKOV MODELS

A hidden Markov model (HMM) is the statistical modeling of a memoryless stochastic process for which the output is observable while the internal state of the model producing this output remains hidden. A more seizable description is to visualize HMMs as directed graphs where each node has associated probabilities for transitioning into other nodes through a directed edge and for emitting certain output (figure 2.6).

To describe these models, we let $K = \{k_1, \ldots, k_n\}$ be the set of *states* (graph nodes) and $B = \{b_1, \ldots, b_m\}$ the alphabet of *observations*. A special state 0 models both beginning and end of a sequence. In addition, we define the *transition matrix* $A \in \mathbb{R}^{n \times n}$ with a_{ij} as the probability of transitioning from state i to state j and the *emission ma-*

1 The descriptions of HMMs, profile HMMs, and profile HMM-sequence comparisons, as well as the used nomenclature are to a large degree based on Durbin *et al.* (1998).

trix $E \in \mathbb{R}^{n \times m}$ with the probability of observing b in state k denoted as $e_k(b)$. To relate these definitions to the aforementioned hidden model states and observable output, we further introduce the sequence of hidden states $\pi = \{\pi_1, \ldots, \pi_t\}$ with timepoints $T \in \mathbb{N}$ and the observed sequence $x = \{x_1, \ldots, x_t\}$.

These definitions allow us to describe HMMs more accurately. The aforementioned *memorylessness* is also called the Markov property. Our model has the Markov property as the conditional probability for the next state is

$$a_{kl} = P(\pi_i = l | \pi_{i-1} = k)$$

and thus solely depends on the current state while being independent of all other states and observations. Similarly the emission probability of b in state k is

$$e_k(b) = P(x_i = b | \pi_i = k),$$

which again only depends on the current state. These two conditional probabilities are represented by the directed edges in the graph representation for both state transitions and emissions (figure 2.6).

The aforementioned definitions indicated that the usefulness of an HMM depends critically on its topology as well as on transition and emission probabilities. Model topology is often decided upon by expert judgment and will concern us more specifically later in the discussion of profile HMMs (see page 18). The approach for determining transition and emission probabilities depends on whether we know the path through the model for the observations in our training dataset. If the path is unknown, the iterative Baum-Welch and Viterbi training methods can be used (Baum, 1972; Durbin *et al.*, 1998). Due to our aforementioned focus on profile HMMs we need not concern ourselves further with these algorithms as in our use case we know both the state path and the observations for all training data (see page 18).

From the known paths, we can directly infer the probability of transitioning from state k to l as the fraction of all outbound transitions

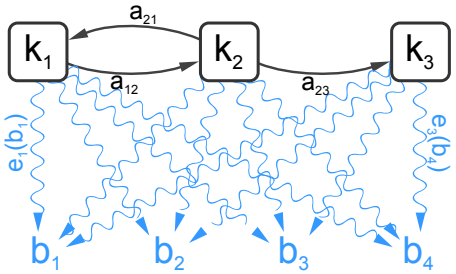


Figure 2.6: Graph representation of a hidden Markov model (HMM) with three states k_i, transition probabilities a_{kl} (arrows), four observations b_j (blue), and emission probabilities $e_k(b)$ (blue waves). Start and end states were omitted, as were most emission probability labels.

of state k that lead to state l. With A_{kl} the number of transitions from k to l in the training dataset, we use

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}. \tag{2.1}$$

Likewise, the emission probability for b in state k is determined based on its number of occurrences $E_k(b)$, which yields

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}. \tag{2.2}$$

Transitions and emissions that are absent from the training dataset result in forbidden transitions and emissions in the HMM. This issue is often encountered with small training datasets or with sets of little diversity and the usual way of avoiding problems is to add pseudocounts to A_{kl} and $E_k(b)$ (Durbin *et al.*, 1998).

As we now have a way of parametrizing models given training data, we can assess properties of observed sequences with respect to the model. The joint probability of an observed sequence x and a state sequence π can be computed by moving along the state sequence and recording the probabilities for transitioning into the next state π_k and for emitting the observed output x_k in that state. Including the probability of starting in state π_1 and ending in end state $0 = \pi_{L+1}$, we arrive at

$$P(x, \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}},$$

which requires knowledge about the sequence of states that we usually lack.

This dilemma can be circumvented if we assume that the most probable path through the model

$$\pi^* = \operatorname*{argmax}_{\pi} P(x, \pi).$$

or the full probability of the sequence given the model

$$P(x) = \sum_{\pi} P(x, \pi)$$

are the properties we are interested in.

The Viterbi algorithm computes π^* through the Viterbi variables $v_k(i)$, which capture the maximal probability of emitting $x_1 \dots x_{i-1}$ from any sequence of states followed by x_i from state k (Viterbi, 1967). These variables can be computed recursively by maximizing the probability of transitioning into state l and emitting x_{i+1} from it after $x_1 \dots x_i$ were emitted along the most probable path, which condenses to

$$v_l(i+1) = e_l(x_{i+1}) \max_k (v_k(i) a_{kl}). \tag{2.3}$$

With the addition of start state 0 (i. e. $v_0(0) = 1$ and $v_k(0) = 0 \forall \{k | k \in K, k \neq 0\}$), this equation is then used in a dynamic programming recursion. While the most probable path is often interesting on its own merits, we can also compute the full probability of emitting the observed sequence by replacing the maximization in equation 2.3 with a sum; this is the *forward algorithm* (Durbin *et al.*, 1998).

PROFILE HIDDEN MARKOV MODELS

The definitions and algorithms previously introduced for HMMs help us in the explanation of profile HMMs, which in sequence bioinformatics are mainly used to represent and compare protein families and to search for sequences related to these families. Like sequence profiles, profile HMMs are derived from MSAs of homologous proteins and benefit from *deep* alignments, i. e. those with a large number of diverse sequences. The main advantage of profile HMMs over sequence profiles are position-specific gap scores versus a uniform gap treatment.

The states of each profile HMM are arranged in a set of layers of identical structure, each consisting of three states for *matches*, *insertions*, and *deletions* (figure 2.7; Durbin *et al.*, 1998). The match state is used to model family consensus residues, whereas the insertion and deletion states represent additional and skipped amino acids relative to the family, respectively.

In the model, match state M_i captures the amino acid frequencies of the ith family consensus residue as emission probabilities and further comprises the probabilities for continuing with the subsequent family consensus residue, starting an insertion or omitting the next consensus residue. With the emission probability for amino acid a

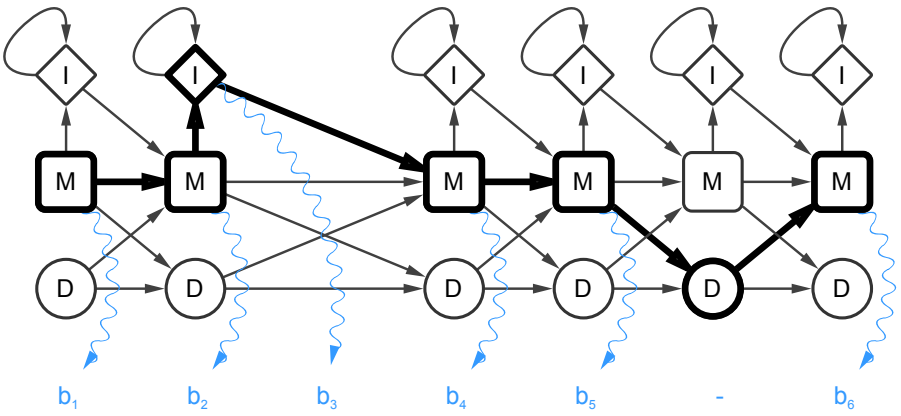


Figure 2.7: Graph representation of a profile HMM with insertion, match, and deletion states in each layer. The path through the model that generated the observations (blue waves and labels) is highlighted. Adapted from Söding (2005).

from state M_j denoted as $e_{M_j}(a)$, an emission contributes to the final score with the log-odds ratio

$$\log \frac{e_{M_j}(a)}{f(a)},$$

where $f(a)$ is the background frequency of amino acid a. The log-odds score S_{LO} of a sequence $x = \{x_1, \ldots, x_L\}$ to be emitted on a path π through the model thus is

$$S_{LO} = \log \frac{P(x|\pi)}{P(x|F)}, \tag{2.4}$$

where $P(x|F) = \prod_{l=1}^{L} f(x_l)$ is the probability to emit the sequence under the null model F, i. e. the background amino acid frequencies.

Insertion state I_j (diamonds in figure 2.7) models one or more residues inserted after consensus residue i is matched. These states are special as they are the only state type in profile HMMs with transitions to themselves. The probability of these loop transitions $I_j \to I_j$ reflects the distribution of insertion lengths encountered after matching family residue j. Assuming that inserted residues follow the background amino acid frequencies, the insert state has a log-odds score (equation 2.4) of 0, irrespective of the length of the insertion. The overall score contribution of a k residues long insertion is

$$\log a_{M_j I_j} + (k-1) \log a_{I_j I_j} + \log a_{I_j M_{j+1}},$$

which means that the position-specific insertion scoring is encoded in the probabilities of transitioning to, within, and away from insertion states.

Deletion states are the third and final state type (circles in figure 2.7) and allow for proper modeling of sequences that lack consensus residues of the family. This is achieved by making the deletion states *silent*, which means that they do not emit any symbol. Every match state M_j can be skipped through an associated deletion state D_j, which is reachable from states M_{j-1} and D_{j-1} in the previous layer. The use of one deletion state for each skipped match makes the deletion scoring position-specific.

To create profile HMMs that represent a protein family, we derive the parameters of our model from an MSA of the family. It is an important decision whether an alignment column is considered part of the family consensus and becomes a match state or whether it is treated as an insertion and should contribute to an insertion state. A rule-of-thumb is to treat columns with less than 50% gaps as matches whereas the others are considered insertions; a different rule useful in its specific context will be discussed at the end of section 2.2.5. Once this decision is made, the transition and emission probabilities of the profile HMM can be derived directly from the MSA using equations 2.1 and 2.2. As with HMMs, pseudocounts must be added before computing these

probabilities to avoid problems with transitions or emissions that are never observed and would thus not be possible in the derived model (Durbin *et al.*, 1998).

Finally, modified versions of the previously introduced algorithms can now be used to efficiently compute the most likely path to generate a sequence or its total probability. As an example, updating the Viterbi variables now specifically only considers the topologically possible inbound transitions for a state instead transitions from all other states. The single dynamic programming matrix is replaced by one matrix for each of the three state types. The Viterbi variable (equation 2.3) in the match matrix, for example, is modified to

$$V_j^M(i) = \log \frac{e_{M_j}(x_i)}{q(x_i)} + max \begin{cases} V_{j-1}^M(i-1) + \log a_{M_{j-1}M_j}, \\ V_{j-1}^I(i-1) + \log a_{I_{j-1}M_j}, \\ V_{j-1}^D(i-1) + \log a_{D_{j-1}M_j}; \end{cases}.$$

The other two matrices are derived analogously except for the lacking log-odds score in the deletion matrix. The initializations need to be slightly altered as well (for details see Durbin *et al.*, 1998).

Finally, the most probable path to generate a sequence can be used to align this sequence to any of the sequences in the MSA from which the profile HMM was derived due to the direct correspondence of alignment columns to match and insertion states. The result of a profile HMM-sequence comparison thus often includes an alignment to a representative sequence of the profile HMM.

PROFILE HMM-SEQUENCE COMPARISON

Two important applications for profile HMMs in sequence bioinformatics are querying a large sequence database with a profile HMM and querying a database of protein family profile HMMs with a sequence. The former application is comparable to the approaches presented in the previous sections and can be considered a more sensitive replacement for searches otherwise performed with BLAST or PSI-BLAST. To this end it is noteworthy that current implementations of profile HMM algorithms are approximately as fast as BLAST and PSI-BLAST on this task, as shown for HMMER3 (Eddy, 2011). The latter application is most prominently featured in Pfam, a web resource offering descriptions, MSAs, profile HMMs, and further details for almost 15000 protein families (Pfam 27.0 of March 2013; Finn *et al.*, 2014). Pfam uses the aforementioned HMMER3 package internally to search its profile HMMs with user queries and shows all families with significant matches to the query sequence.

2.2.5 *Pairwise profile HMM comparison*

We previously showed that profile HMMs capture more of the information encoded in MSAs than sequence profiles. We also mentioned

that the comparison of sequence profiles to each other instead of to sequences improves the remote homology detection sensitivity (section 2.2.3). It thus seems natural to extend on profile HMM-sequence comparisons by performing pairwise profile HMM comparisons and indeed the currently most sensitive approach for remote homology detection, HHsearch, implements this idea (Söding, 2005; Söding *et al.*, 2005).[2]

For HHsearch, the log-odds score (equation 2.4) was generalized to a log-sum-of-odds (LSO) score that reflects the co-emission probability of two profile HMMs. The basis of the LSO score is the column score

$$S_{aa}(q_i, p_j) = \log \sum_{a=1}^{20} \frac{q_i(a), p_j(a)}{f(a)}, \tag{2.5}$$

which compares the amino acid distributions of the match states in layers i and j of profile HMMs q and p, respectively. The column score has properties similar to the log-odds ratio in profile HMM-sequence comparisons, e. g., it vanished for insertions and columns that follow the background amino acid frequencies (Söding, 2005). With the column score, the path P through the two profile HMMs, and P_{tr} as the total probability of transitioning P, the LSO score can be derived as

$$S_{LSO} = \log \sum_{g:X_g Y_g = MM} S_{aa}\left(q_{i(g)}, p_{j(g)}\right) + \log P_{tr}, \tag{2.6}$$

with X and $Y \in \{M, I, D\}$ state types in q and p, respectively, G the number of columns in the alignment of the two profile HMMs, X_g and Y_g the states in the gth column of the alignment and $i(g)$ and $j(g)$ the layers of q and p corresponding to g, respectively.

To complete the description of pairwise profile HMM alignments, gaps are introduced with identical meaning as in sequence alignments (section 2.2.1). Gaps are not modeled as states, however they are be considered during the dynamic programming used to optimize the LSO score.

Certain transitions are not allowed and the alignment is limited to pair states $XY \in \{MM, MI, IM, DG, GD\}$, where G denotes a gap. In this notation, MI is the alignment of a match state from profile HMM q to a insertion state in p. A modified Viterbi algorithm is used to find the pairwise profile HMM alignment with maximal LSO score by recursively computing individual dynamic programming matrices for each pair state XY (Söding, 2005). Finally, this results in a score for the alignment of two profile HMMs.

Instead of comparing profile HMMs that represent protein families, it can be useful to compare profile HMMs that represent particular proteins. The assignment of MSA columns to match and insertion states should then be changed to simply declaring columns with residues in

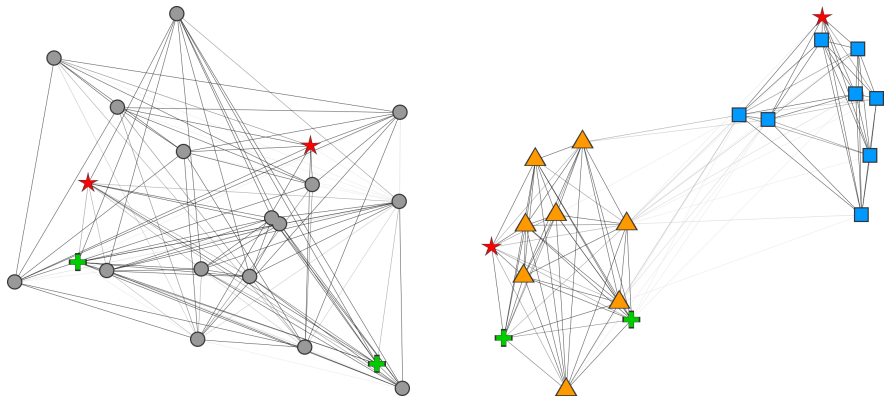
2 The description of pairwise profile HMM comparisons is based on Söding (2005).

the query sequence to matches. Profile HMMs derived with this new rule allow for comparisons tailored to single sequences that are superior in reflecting sequence peculiarities instead of a family consensus.

2.3 CLUSTER ANALYSIS OF SEQUENCES

The cluster analysis of sequences (CLANS) software visualizes the pairwise similarities of a set of protein sequences as a graph, which we often refer to as a *cluster map* (Frickey and Lupas, 2004). We alter the terminology of graphs for cluster maps and refer to nodes as points or sequences and to edges as connections or similarities. CLANS represents protein sequences as symbols in a 2D space and their pairwise similarities as connecting lines.[3] Random starting positions are then optimized with the iterative force-directed graph layout algorithm by Fruchterman and Reingold (1991, figure 2.8).

This algorithm optimizes a tripartite force equation. The first term models pairwise repulsions between points and fades gradually with their distance. The second term contributes a small force towards the coordinate system origin to keep unconnected subgraphs from drifting apart due to the first term. The final term adds attractive forces between connected points. By default CLANS uses the negative logarithm of BLAST p-values as attractive forces (section 2.2.2). While the



A B

Figure 2.8: A) CLANS graph after random initialization of positions. Sequences are represented as icons and pairwise similarities as lines; the darker a line the more similar the sequences. The green crosses were chosen as an example of connected points whereas the red stars are unconnected to each other. B) The data from A after layouting. Icons were changed to reflect cluster memberships in the orange triangle and blue square cluster; the green crosses and red stars were retained.

3 CLANS offers 3D layouting, however the dimensionality reduction necessary for a printable image is not trivial and effectively leads to a 2D layout. Therefore, all graphs presented in this work were optimized in 2D mode.

first two force terms are relevant to all (pairs of) points, the third term only creates attraction where similarities above a user-chosen threshold exist. In each iteration of the algorithm, the total force effecting each sequence is computed as the sum of the three terms and the sequence position is altered according to it. As convergence criterion, we observe whether topologically relevant changes still occur over a number of rounds once the layout has visually stabilized. To avoid local minima, the clustering should be computed from multiple random initializations.

Obtaining a reasonable graph layout is only the first step and is followed by extensive semi-automatic and manual analyzes of features of the dataset, e. g., family membership and domain arrangement of a sequence. These features are visualized on the map by different symbol colors and shapes. The members of one sequence cluster are visually grouped through identical symbol properties.

2.4 PAIRWISE STRUCTURE COMPARISON

Besides the sequence-based comparisons introduced in the previous sections, it can be useful to assess the structural similarity of two proteins. While homology cannot be established by structural similarity alone (section 2.1.3), many applications require or benefit from this information. . In this work, we used structure comparison algorithms, e. g., to create reliable alignments of remote homologs with known structure.

Most algorithms compare the positions of corresponding residues often represented by their C_α coordinates unless side-chains are of special interest. A pairwise residue assignment for two proteins can be derived from the match columns of their pairwise sequence alignment (section 2.2.1). Depending on the algorithm, the initial alignment might be optimized in conjunction with the similarity score.

The most popular protein structure similarity score is the root-mean-square deviation (RMSD), given as

$$\mathrm{RMSD}(v, w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2},$$

where v and w are sets of n corresponding positions in two superimposed proteins and $\|v_i - w_i\|^2$ denotes the squared length of the vector from v_i to w_i. Given the corresponding positions, a superimposition with minimal RMSD is found by translating the protein centroids onto each other and then computing the optimal rotation of one protein with the Kabsch algorithm (Kabsch, 1976, 1978).

As evident from the definition, outliers, i. e. large distances between corresponding positions, can become a predominant part of the score, limiting its usefulness in general and for superimposition. Further, if

the alignment is optimized in parallel to the RMSD, finding a reasonable balance between the number of aligned residues and the score is difficult (Zhang and Skolnick, 2004).

Other scoring functions tried to address these issues and within this work we used the TM-score, which was introduced along with the TM-align algorithm that optimizes it efficiently (Zhang and Skolnick, 2004, 2005). The TM-score is defined as

$$\text{TM-score} = \max\left[\frac{1}{L_\text{N}}\sum_{i}^{L_\text{ali}}\frac{1}{1+\left(\frac{d_i}{d_0(L_\text{N})}\right)^2}\right],$$

where the normalization value L_N is one of the protein lengths (or their mean), L_ali is the number of aligned residues, and d_i is the distance of the C_α atoms of aligned residue pair i. The term $d_0(L_\text{N}) = 1.24\sqrt[3]{L_\text{N}-15}-1.8$ was derived to make the average TM-score of random structure pairs independent of the protein size. It represents the expected average distance of randomly paired residues in structures of length L_N. The TM-score is in the interval $]0,1]$ and perfectly identical structures have a value of 1 whereas random pairs of structures are scored ~0.17 (Xu and Zhang, 2010).

TM-align optimizes this score by estimating several initial alignments and then optimizing all of them with iterated dynamic programming rounds. Before each iteration, one protein is rotated optimally with respect to the TM-score given the current alignment. In practice, the algorithm converges towards a stable alignment quickly, aligning even large proteins within seconds.

2.5 STRUCTURE-AIDED HOMOLOGY VALIDATION

As homologies become more remote and the number of residues that can be compared decreases, it becomes progressively harder to establish statistically significant similarity between sequences over the background. In such cases it would be beneficial to include structural information into the comparisons, because, even though prone to convergence, structures diverge more slowly than sequences (section 2.1.3). A method to do this was recently introduced in order to establish cases of distant homology (Remmert *et al.*, 2010).

Its rationale is that homologs were almost identical in sequence and structure when they started to diverge from their common ancestor. Over time, these proteins accumulated differences, resulting in progressively lower similarities both in sequence and, more slowly, also in structure. Due to the continuity of this process, we expect to see a correlated decrease in sequence and structure similarity for homologous proteins. Analogs should have varying degrees of structural similarity, mostly independent of sequence similarity, and sequence

similarities should generally be lower. Sequence and structural similarity scores of analogs are expected to be uncorrelated.

It is conceivable that specific local structures might restrict the possible amino acids at one or more positions of the protein, leading to a similar correlation between structure and sequence similarity. A test of this possibility in a study on the origin of outer-membrane β-barrels did not uncover such correlation (Remmert *et al.*, 2010), as expected from the observation that domains are multiply convergent at the supersecondary structure level without an accompanying increase in sequence similarity.

2.5.1 *Calculation*

First, each query-template pair is aligned with TM-align and the query length-normalized TM-score is obtained, which we use as our structural similarity score (section 2.4). Next, the sequence similarity is computed based on the TM-align alignment by comparing the corresponding columns of query and template profile HMMs with HHalign, the HHsearch scoring procedure (section 2.2.5). The score HHalign returns is normalized by the number of aligned residues. For the sake of simplicity, we call these sequence similarities 'HHalign scores'. Finally, SciPy (Jones *et al.*, 2001) is used to calculate the correlation between TM-score and HHalign scores for curated subsets of the query and template datasets.

2.5.2 *Visualization*

The data points are plotted with TM-score on the horizontal and HHalign score on the vertical axis. Additionally, we show a regression line along with ellipses representing the first three standard deviations around the mean on top of the data points (e. g., figure 4.7).

2.5.3 *Significance*

To derive the statistical significance, a linear dependency between TM-score and HHalign scores is assumed. For each set of comparisons (e. g., the set of scores of all comparisons of inositol-requiring enzyme 1 luminal domain (IRE1-LD) queries against PQQ motif β-propeller templates), a linear regression is computed with SciPy along with a t-test with the null hypothesis that the slope is zero. In other words, the existence of a significant relationship between TM-score and HHalign is assessed. With a significance level of $1e^{-3}$ as threshold the correlation values in this manuscript imply significant relationships unless otherwise noted.

3

THE TUBULAR LIPID-BINDING DOMAIN SUPERFAMILY

This chapter includes content from the following publications.
Re-use permissions have been granted by the copyright holders.

Kopec K.O., Alva V., and Lupas A.N.
Bioinformatics of the TULIP domain superfamily.
Biochemical Society transactions, *39*(4): 1033–1038, 2011.

Kopec K.O., Alva V., and Lupas A.N.
Homology of SMP domains to the TULIP superfamily of lipid-binding proteins provides a structural basis for lipid exchange between ER and mitochondria.
Bioinformatics, *26*(16): 1927–1931, 2010.

3.1 INTRODUCTION

Most eukaryotes comprise mitochondria, an organelle of endosymbiotic origin. Besides providing the majority of adenosine triphosphate (ATP) molecules, on which cells depend critically as energy supply, they are also involved in apoptosis, amino acid and lipid metabolism, iron-sulfur cluster assembly, and the regulation of calcium levels within the cell (Lill and Kispal, 2000; McBride *et al.*, 2006). Most biopolymers used in these processes are not produced in the mitochondria themselves but instead imported. A prominent example are phospholipids crucial for the mitochondrial membranes that must be imported from the endoplasmic reticulum (ER). An interesting case among the phospholipids is phosphatidylcholine. For its generation, mitochondria import phosphatidylserine from the ER and decarboxylate it to phosphatidylethanolamine, which is then exported again to the ER and methylated to phosphatidylcholine (Voelker, 2003). It is unclear, how this exchange of phospholipids between ER and mitochondria proceeds.

While most organelles use vesicles to transfer phospholipids, mitochondria potentially use sites of direct contact to the ER for the exchange (Achleitner *et al.*, 1999; Voelker, 2003). In *Saccharomyces cerevisiae*, the endoplasmic reticulum-mitochondria encounter structure (ERMES) has been found to tether ER and mitochondria together and allow for efficient phospholipid transport between the two organelles (Kornmann *et al.*, 2009). It however remains elusive whether the effi-

ciency impact of ERMES results from an active role of it in the transfer or its tethering that allows other proteins to perform the transfer.

Four proteins constitute the ERMES complex (figure 3.1). One protein is located in the outer mitochondrial membrane (Mdm10), the second one putatively also in the membrane (Mdm34, also called Mmm2), the third one is ER-resident (Mmm1), and the fourth one is found in the cytosol (Mdm12; Kornmann *et al.*, 2009). Interestingly, two of the proteins (Mmm1 and Mdm12) contain a synaptotagmin-like, mitochondrial, and lipid-binding proteins (SMP) domain. This domain has so far only been described bioinformatically and proteins comprising it were grouped as C2 domain synaptotagmin-like, PH domain-containing HT-008, PDZK8, and mitochondrial proteins (Lee and Hong, 2006). All previously known SMP domain-containing proteins are from eukaryotes and are membrane-associated, however their function is mostly unknown. Given this abundance of SMP domains and the lack of knowledge surrounding them, we decided to investigate the ERMES complex in detail using state-of-the-art bioinformatic methods.

Our main goal was to establish a comprehensive account of the distribution and evolution of SMP domains. In addition, we were interested in homologs with known structure or function that could provide information about the role of SMP domains in the ERMES complex and possibly about the function of the complex itself.

3.2 MATERIALS AND METHODS

All sequence similarity searches were carried out in the MPI Bioinformatics Toolkit (`http://toolkit.tuebingen.mpg.de`; Biegert *et al.*, 2006) using HHpred (version 1.6; Söding *et al.*, 2005) and HHsenser (no version number available; Söding *et al.*, 2006) with default settings.

3.2.1 *SMP domain detection*

The initial searches for SMP domains with HHsearch used the SMP domain from Mmm1 (accession details in figure 3.2; Lee and Hong,

Figure 3.1: Domain organization of the four ERMES complex proteins. All ERMES proteins, except the mitochondrial outer membrane protein Mdm10, contain an SMP domain. The SMP domain in Mdm34 was discovered in this work.

2006) as query against a custom database. This database comprised PDB70 (as available on the 15th of April 2010) and the genomes of phylogenetically diverse organisms (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, and *Saccharomyces cerevisiae*). Representatives of the four SMP domain-containing groups from the aforementioned organisms were later chosen as seeds for the searches in figure 3.2 and figure 3.4, based on their presence in the core of their respective clusters in the sequence cluster map (figure 3.6).

3.2.2 *Cluster map dataset*

To identify sequences for cluster analysis, we searched the nr database at NCBI (as available on the 15th of April 2010) for homologs of several query proteins using HHsenser (Söding *et al.*, 2006). HHsenser combines PSI-BLAST (section 2.2.3) searches from various intermediate proteins with an additional back validation for improved sensitivity and specificity during the establishment of transitive homologies. The first search uses the query protein and the alignment of its high-scoring matches are used to draft an initial family profile. Low-scoring matches are used as query in additional PSI-BLAST searches for which the results are back validated. To this end, the alignment of matches from the current search is converted to a profile HMM and then compared to the family profile using pairwise profile HMM comparisons (section 2.2.5).

If the back validation results in a low score, the query and matches of the current search are discarded. In case of a mediocre score, the matches are added to a set of potential homologs (the *permissive* set), whereas high back validation scores additionally lead to an inclusion of these sequences into the family profile (the *strict* set). The stringent threshold of the strict set ensures that only homologs are added to this set, whereas the permissive set is prone to false positives that require manual curation.

For our dataset, we used the SMP domain from the yeast protein Mmm1 (accession details in figure 3.2), the N-terminal domain of human CETP (2OBD:16–206), the Takeout 1 protein from *Epiphyas postvittana* (3E8T), and the dust mite allergen Der p 7 (3H4Z) as queries. We pooled the permissive sets returned by HHsenser to obtain 2033 sequences, which we clustered by their pairwise BLAST (version 2.2.22) p-values in CLANS (section 2.3; no version number available; Frickey and Lupas, 2004). The p-values were obtained by running BLAST from the CLANS command line interface. Clustering was done to equilibrium in 2D at a p-value cutoff of $1e^{-4}$ using default settings.

3.2.3 *Structure-aided multiple sequence alignment*

The alignment in figure 3.5 was generated by a three-step approach. First, a multiple alignment of SMP sequences was obtained using HHpred in local maximum accuracy alignment mode. Second, a sequence alignment of TULIP domain structures was derived from a multiple structure superimposition calculated using MAMMOTH-mult (no version number available; Lupyan *et al.*, 2005). In the final step, these two alignments were merged manually using as guide an alignment between 1EWF and Mmm1 obtained with HHpred.

3.3 RESULTS

3.3.1 *Member identification*

We systematically searched for SMP domains using the instance in Mmm1 as query (section 3.2.1, figure 3.2). The top matches were to proteins known to contain an SMP domain, but the search further revealed that the ERMES protein Mdm34 also has an SMP domain. In contrast to previous reports of Mdm34 being an integral outer membrane protein (OMP), our analysis of these proteins did not show any signs of a membrane insertion sequence motif (Youngman *et al.*, 2004). We confirmed the detection of an SMP domain in Mdm34 with additional searches using other SMP domains as query. Thus, ERMES indeed comprises not two but three SMP domains.

Interestingly, these searches also returned matches to proteins from the bactericidal/permeability-increasing proteins (BPI)-like family. BPI itself and cholesteryl ester transfer protein (CETP) were among these matches, both of which have known structures (1EWF and 2OBD), as well as lipopolysaccharide-binding protein (LPSBP), lipid-binding serum glycoprotein (LBSGP), phospholipid transfer protein (PLTP), and long (LPLUNC) and short (SPLUNC) paralogs of palate, lung, and nasal epithelium carcinoma-associated protein (PLUNC). Many of these proteins are involved in binding lipids, e.g. CETP helps the transfer of lipids between different lipoproteins (Qiu *et al.*, 2007).

The two aforementioned proteins BPI and CETP are structurally similar; both comprise two tandem domains of the same fold oriented head-to-head. Each domain wraps a α-helix in a highly curved antiparallel β-sheet. Within the BPI-like family, the only exception from this structural template is SPLUNC, which only contains one such domain, whereas LPLUNC comprises two (Bingle and Craven, 2002). The two domains are structurally highly similar, yet their sequence identity is very low ($\leq 15\%$) and indeed sequence searches with the N-terminal domain as query do not match the C-terminal domain and vice versa. The only match made when searching with the C-terminal domain is the Aha1 protein, which adopts the same fold but is

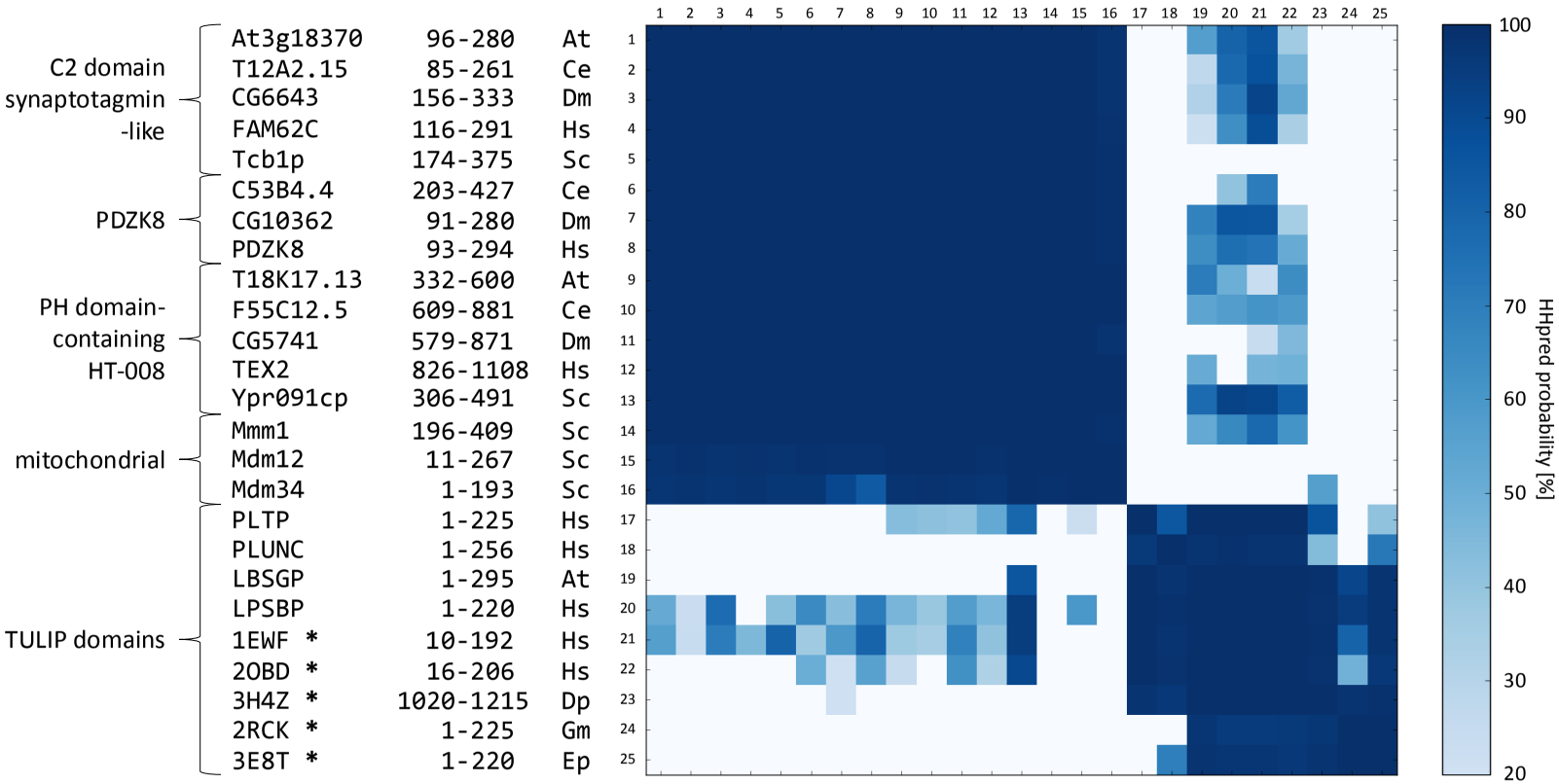


Figure 3.2: Pairwise profile HMM comparison of SMP and TULIP domains. Representatives of the four SMP domain-containing groups and of TULIP domains were chosen from *Arabidopsis thaliana* (At), *Caenorhabditis elegans* (Ce), *Drosophila melanogaster* (Dm), *Dermatophagoides pteronyssinus* (Dp), *Epiphyas postvittana* (Ep), *Galleria mellonella* (Gm), *Homo sapiens* (Hs), and *Saccharomyces cerevisiae* (Sc). Group and protein names of these representatives, along with domain boundaries and names of source species are indicated (from left to right). HHsearch was used to perform pairwise profile HMM comparisons between them. Cell color indicates the HHsearch probability of the match as depicted in the scale on the right; probabilities $\leq 20\%$ are shown as white cells. Proteins with known structures are marked with an asterisk and their PDB accession is provided instead of a name. The abbreviations of the remaining TULIP domains are explained in the text.

classified into a separate superfamily in the SCOP database—d.83.2 for Aha1 and d.83.1 for the C-terminal domain (Murzin *et al.*, 1995). Even though it cannot be detected on the protein sequence level, the N- and C-terminal domains are thought to derive from a common ancestor due to their remarkably high structural similarity and structural classification of proteins (SCOP) considers them members of the same family (d.83.1.1; Kleiger *et al.*, 2000).

The low sequence similarity between the N- and C-terminal domains is also reflected by the results of various HHpred searches with SMP domains as query, where only matches to the N-terminal domain were found. Reciprocal searches with both domains of BPI-like proteins support this finding. The matches between SMP domains and BPI-like N-terminal domains have significant and high probabilities, from which we infer that these two group are homologous. We thus assume that they adopt the same fold and have similar lipid-binding properties.

Among the results of searches with the N-terminal domain of BPI-like proteins were three further proteins with known structures. Dust mite allergen Der p7 (3H4Z), a juvenile hormone-binding protein from *Galleria mellonella* (JHBP, 2RCK), and a Takeout 1 protein from *Epiphyas postvittana* (3E8T) are arthropods proteins involved in binding hydrophobic ligands. In contrast to the tandem domain BPI-like proteins, these proteins comprise only one domain of this fold, a homolog of the N-terminal domain (figure 3.3. This relationship between these two protein families was noted before (Hamiaux *et al.*, 2009; Kolodziejczyk *et al.*, 2008; Mueller *et al.*, 2010). Due to the high degree of sequence and structure similarity, we consider the arthropod and BPI-like proteins families of the same superfamily. We named this superfamily tubular lipid-binding proteins (TULIP).

3.3.2 *Fold prediction*

To substantiate the relationship between SMP and TULIP domains so far established using HHpred (section 3.3.1), we queried four fold prediction servers with a set of 16 representative SMP domain sequences (figure 3.4; analyses performed in April and Mai 2010). We used Phyre, MUSTER, and MULTICOM, as well as the metaserver I-TASSER (table 3.1), all of which were top-ranked in critical assessment of protein structure prediction (CASP) experiment 8 (Kelley and Sternberg, 2009; Wang *et al.*, 2010; Wu and Zhang, 2008; Roy *et al.*, 2010; Kryshtafovych *et al.*, 2009). I-TASSER was the top server in CASP 7 and 8 (Zhang, 2007, 2009). All four servers predicted a TULIP domain as top match in the majority of the 16 cases (Phyre 15; MULTICOM 11; MUSTER 13; I-TASSER 10). Except for MULTICOM, which only shows the top result, the servers returned three to ten matches in a ranked list. Considering the top 3 positions in these lists, all but one query SMP domain

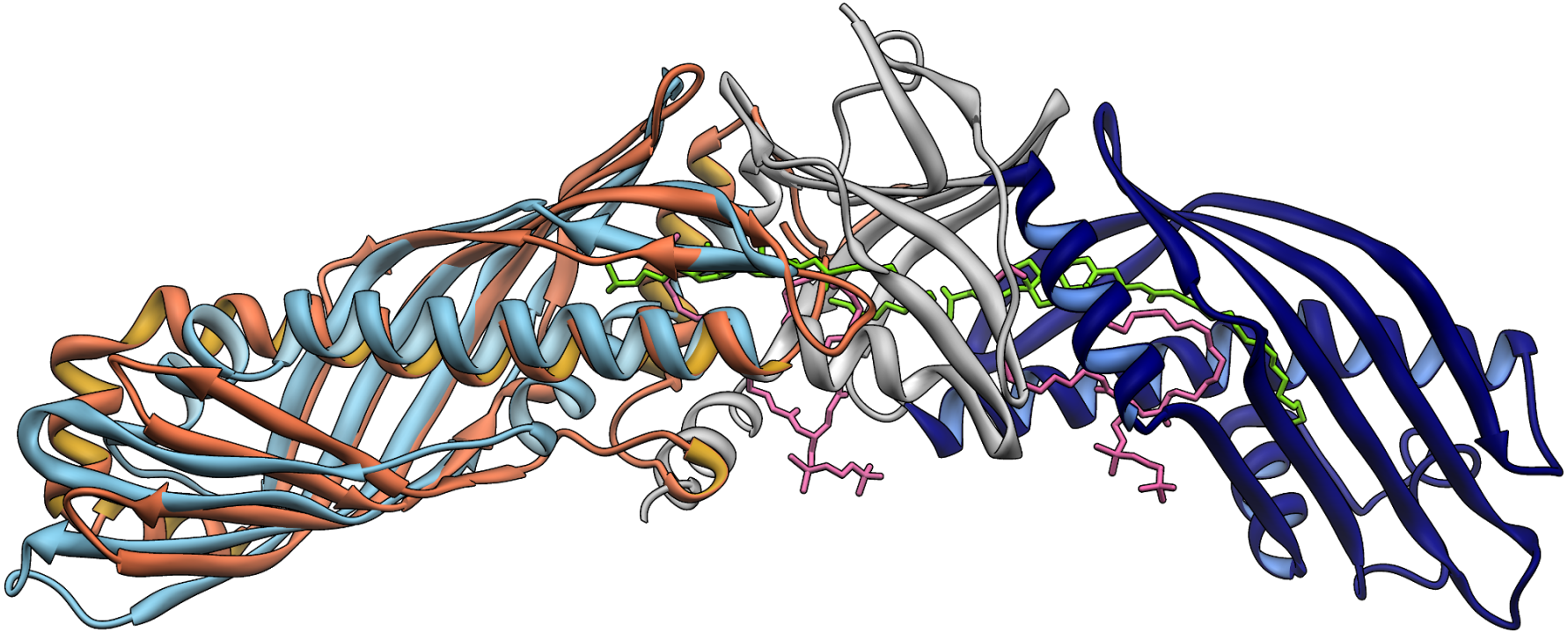


Figure 3.3: Structural superposition of CETP and JHBP. JHBP (2RCK) is shown in orange, the N-terminal part of CETP (2OBD:20–196) in cyan, and the C-terminal part of CETP (266–436) in blue. The collar region of CETP is shown in gray. The bound phospholipids and cholesteryl esters of CETP are shown in magenta and green, respectively.

have matches to a TULIP domain; the query without this match varies between the methods (table 3.1). Overall, the four servers thus confirmed the connection between SMP and TULIP domains.

The structures matched by the fold prediction servers contained BPI- and Takeout-like proteins. We thus created a structure-aided multiple sequence alignment of SMP domains and these TULIP domains (figure 3.5). The alignment showed that length, (predicted) secondary structure, and hydrophobicity patterns are similar, which explains the fold prediction server matches. We found no conserved sequence motifs, which is in agreement with previous reports of lacking motifs even within these families (Beamer *et al.*, 1997; Kolodziejczyk *et al.*, 2008).

3.3.3 *Cluster map*

To assess the similarity between different TULIP superfamily domains, we created a cluster map (section 3.2.2, figure 3.6). In the cluster map, we found confirmations for the proposed homologies between SMP, BPI, and Takeout-like domain families as they were found in clearly separated yet still connected regions of the cluster map. While Lee and Hong (2006) describe several SMP domain-containing proteins and we discussed another one here, Mdm34, we found that there is another cluster in the SMP group, which comprises the uncharacterized transmembrane 24 proteins. We also found several groups of BPI-like proteins, among them the expression site-associated gene 5 (ESAG5) proteins which were indicated to be BPI homologs (Barker *et al.*, 2008). We also found a case of domain tandems with high sequence similarity in arthropod allergens, which is indicative of a du-

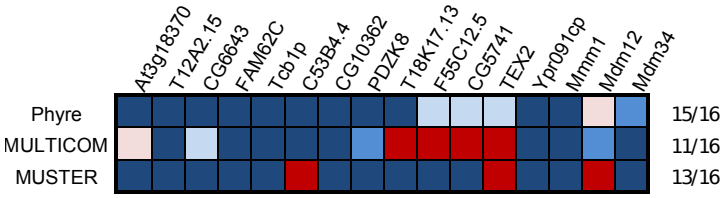


Figure 3.4: Summarized results of SMP domain fold predictions. The highest scoring PDB matches for the 16 representative SMP sequences in figure 3.2 were collected with three of the top-scoring prediction servers in CASP8. Top matches to TULIP domains are shown in blue and to any other structure in red. The color saturation is scaled linearly between the maximum and minimum scores by the respective method. The value ranges corresponding to pale (low confidence), medium, and dark (high confidence) saturation are: Phyre estimated precision 0–33, 34–66, and 67–100, MULTICOM E-value 7.4–5, 5.1–2.6, 2.5–0, MUSTER Z-score 0–1.8, 1.9–3.5, 3.6–5.3. The number of matches to TULIP domains against the total is shown to the right.

A

Phyre	1st		2nd		3rd	
At3g18370	1ewf	90.0	1bp1	90.0	2obd	85.0
T12A2.15	1bp1	90.0	1ewf	90.0	2obd	75.0
CG6643	1bp1	90.0	1ewf	90.0	2obd	75.0
FAM62C	1bp1	85.0	1ewf	85.0	2obd	65.0
Tcb1p	1bp1	85.0	1ewf	80.0	2obd	60.0
C53B4.4	1bp1	80.0	1ewf	75.0	2obd	65.0
CG10362	1bp1	75.0	1ewf	65.0	2obd	30.0
PDZK8	1bp1	80.0	1ewf	70.0	2obd	60.0
T18K17.13	1bp1	85.0	2obd	35.0	2z63	15.0
F55C12.5	2obd	20.0	1ewf	20.0	1bp1	20.0
CG5741	1bp1	20.0	1ewf	5.0	2r1c	5.0
TEX2	1bp1	25.0	1ewf	15.0	1t11	5.0
Ypr091cp	1bp1	95.0	1ewf	95.0	2obd	90.0
Mmm1	1bp1	85.0	1ewf	85.0	2obd	45.0
Mdm12	1z87	5.0	2qxl	0.0	2dyk	0.0
Mdm34	1ewf	55.0	1bp1	40.0	2gia	30.0
	15/16		15/16		11/16	

B

MUSTER	1st		2nd		3rd		4th		5th		6th		7th		8th		9th		10th	
At3g18370	3e8w	5.3	2co	4.1	2hzq	3.9	1gka	3.9	1w5d	3.8	1bp1	3.7	2wke	3.7	2k23	3.6	2rck	3.6	1ew3	3.6
T12A2.15	3e8w	5.3	3dze	3.6	1aqp	3.6	2aco	3.6	2rck	3.6	3f5b	3.5	2ns9	3.5	3fiq	3.5	1tlj	3.5	2r55	3.4
CG6643	3e8w	5.1	2rck	3.9	1tlj	3.6	1aqb	3.5	2aco	3.4	3dze	3.4	3f5b	3.4	2bgc	3.3	2w7q	3.3	1ttw	3.3
FAM62C	3e8w	5.1	2rck	3.8	3dze	3.7	1tlj	3.6	2co8	3.5	2aco	3.5	2bgc	3.4	2fiq	3.4	2hhp	3.4	2cnw	3.4
Tcb1p	3e8w	5.0	2rck	3.6	2w7q	3.6	1aqb	3.6	1tlj	3.5	2r55	3.4	2w1z	3.4	3fos	3.2	2zpd	3.2	3bzy	3.2
C53B4.4	2pn5	4.3	1bp1	4.2	2ipi	4.2	2pn5	4.1	2wdw	4.0	1bp1	3.9	3jq0	3.8	3jq1	3.7	2ive	3.7	2n6f	3.6
CG10362	3e8w	5.0	3kf8	3.9	2aco	3.7	1gm6	3.6	1ew3	3.5	2k23	3.5	2hzq	3.5	2rck	3.4	1gka	3.4	2wke	3.4
PDZK8	3e8w	4.9	2qgu	3.9	2rck	3.7	3fos	3.6	2wke	3.6	2k23	3.6	3bw8	3.6	1w5d	3.5	1ew3	3.5	1ngl	3.5
T18K17.13	1bp1	4.5	3d2d	4.0	2wyh	3.8	3bnm	3.8	1pn0	3.7	3boa	3.7	3f8u	3.7	1h2x	3.7	1z68	3.6	3ida	3.5
F55C12.5	1bp1	4.5	3d2d	3.9	1sez	3.9	3d2d	3.8	3bg7	3.8	2ipi	3.8	1yf2	3.7	3ida	3.7	2z64	3.6	1pn0	3.6
CG5741	1bp1	4.7	3bg7	4.0	3a0f	3.9	2ipi	3.9	1xfd	3.9	3d2d	3.9	2zun	3.9	2bkl	3.8	3fsn	3.8	3bg7	3.8
TEX2	1h54	3.9	2ipi	3.9	2z64	3.8	2pn5	3.8	3bg7	3.8	1sez	3.7	1obr	3.6	1r1h	3.6	2vfv	3.6	2b39	3.6
Ypr091cp	1bp1	5.1	1bp1	4.7	2r4n	3.8	2vvm	3.6	3k7m	3.5	2nvw	3.5	3e8w	3.5	2ipi	3.5	1gos	3.5	1n6f	3.4
Mmm1	1bp1	3.9	2pn5	3.9	1q21	3.9	2g47	3.8	3e8w	3.8	2b39	3.8	1h2x	3.6	3boa	3.6	3i6d	3.6	3lvt	3.6
Mdm12	3lvt	4.2	1bp1	4.2	2wyh	4.2	3b2s	4.0	2wdw	4.0	2e1v	4.0	1sez	3.9	2x0l	3.8	3chn	3.8	2pn5	3.8
Mdm34	3e8w	4.4	2obd	3.6	3bdl	3.4	1koo	3.2	1yqf	3.2	2dla	3.2	2art	3.2	1zs7	3.1	3fxi	3.1	2rck	3.1
	13/16		7/16		1/16		0/16		2/16		2/16		1/16		1/16		1/16		1/16	

C

MULTICOM	1st	
At3g18370	1ihm	7.4
T12A2.15	2obd	0.8
CG6643	2obd	7.1
FAM62C	1ewf	0.5
Tcb1p	2obd	0.6
C53B4.4	1ewf	2.4
CG10362	2obd	0.1
PDZK8	1ewf	4.6
T18K17.13	1wfj	0.0
F55C12.5	1rlw	0.0
CG5741	3jzy	0.0
TEX2	1wfj	0.0
Ypr091cp	2obd	0.1
Mmm1	2obd	0.3
Mdm12	2obd	3.3
Mdm34	1ewf	0.3
	11/16	

D

I-TASSER	1st		2nd		3rd		4th		5th		6th		7th		8th		9th		10th	
At3g18370	3e8w	0.9	1ewf	0.8	3ihg	0.8	2rck	1.1	1hqp	1.1	1ewf	0.7	2pgc	0.6	1ewf	0.5	3e8t	0.4	2aco	0.7
T12A2.15	2ns9	1.0	2vmf	1.3	2bgc	1.3	3e8w	0.9	1ewf	0.9	1ewf	0.7	2rck	1.0	1qbe	1.1	2q9k	0.6	2rhk	0.4
CG6643	2rck	1.0	2bgc	1.2	3e8w	0.9	1ewf	1.0	1ewf	0.6	2ns9	1.0	1qbe	1.1	1ewf	0.5	1erva	0.4	1ewf	0.5
FAM62C	2rck	1.0	2bgc	1.2	2vmf	1.2	3e8w	0.9	1ewf	0.8	3ihg	0.6	3e8t	1.0	1hqp	1.1	1ewf	0.4	1b8w	0.4
Tcb1p	3e8w	0.9	1ewf	0.3	3ihg	0.7	2rck	1.0	3czj	1.0	1et1	0.7	2vli	0.5	1ewf	0.4	3e8t	0.4	2rck	0.6
C53B4.4	2pn5	0.7	1ewf	0.7	1ewf	0.6	1ii7	1.0	1epb	0.9	3lml	0.5	2vv5	0.5	2obd	0.5	1ewf	0.4	1bp1	0.7
CG10362	3e8w	0.9	1ewf	1.0	1ewf	0.8	2pcs	1.0	1epb	1.5	1ewf	0.8	1i50	0.5	1ewf	0.8	1sdo	0.4	3kf8	0.7
PDZK8	3e8w	0.8	1ewf	0.8	1ewf	0.7	1vhm	1.0	1ew3	1.4	1ewf	0.5	2pcs	0.5	1ewf	0.6	3c8c	0.3	2qgu	0.7
T18K17.13	1bp1	0.8	2obd	0.8	2z7x	0.5	1vho	1.2	1vqw	0.8	1yyc	0.5	1ekg	0.6	1ewf	0.4	2wyh	0.3	3d2d	0.7
F55C12.5	3bg7	0.7	2obd	0.7	2wjs	0.5	1vho	1.3	2tma	0.8	1yyc	0.4	1ekg	0.3	1ewf	0.6	3egc	0.6	3d2d	0.7
CG5741	1bp1	0.8	1ewf	0.6	2hzp	0.6	1io1	1.4	2tma	1.0	2obd	0.4	1ekg	0.5	1ewf	0.5	3cfx	0.3	3bg7	0.7
TEX2	1h54	0.7	1ewf	0.6	1yrg	0.5	1ii7	1.3	1vqw	0.9	1wwp	0.6	1ois	0.3	1yyc	0.6	2fge	0.3	2ipi	0.7
Ypr091cp	2o48	1.0	1bp1	0.9	1ewf	0.9	1ewf	0.7	2rck	1.0	1ew3	1.0	1ewf	0.7	2pcs	0.4	1ewf	0.6	3ebg	0.4
Mmm1	1bp1	0.7	1ewf	0.7	1ewf	0.6	1i50	1.0	2vmf	1.0	1ewf	0.7	1y66	0.4	1ewf	0.7	3e8t	0.4	2pn5	0.7
Mdm12	3lvt	0.7	1ewf	0.4	1ziw	0.6	1i50	1.3	1h0g	1.1	2ov7	0.7	1cdl	0.4	1ewf	0.4	3c8c	0.3	1bp1	0.7
Mdm34	3e8w	0.8	1itw	0.6	1h6u	0.6	1ewf	1.1	1yqf	1.1	1ewf	0.7	2dsm	0.6	1ewf	0.5	3e8t	0.4	2obd	0.6
	10/16		12/16		6/16		7/16		4/16		7/16		3/16		12/16		7/16		5/16	

Table 3.1: Detailed results of SMP domain fold predictions using the A) Phyre, B) MUSTER, C) MULTICOM, and D) I-TASSER server. All results returned by each server are shown. Each row is the result for one query (see figure 3.2 and table 3.2 for accession details). Each rank comprises two columns: the match PDB identifier (green for matches to TULIP domains, red else) and its score. Scores are different per server and are color coded by a red-to-green gradient corresponding to increasing confidence (A: estimated probability, B: E-value, C and D: Z-scores). The last row summarizes the number of matches to TULIP domain structures at each rank with colors red (0–4 matches), orange (5–8), yellow (9–12), and green (13–16).

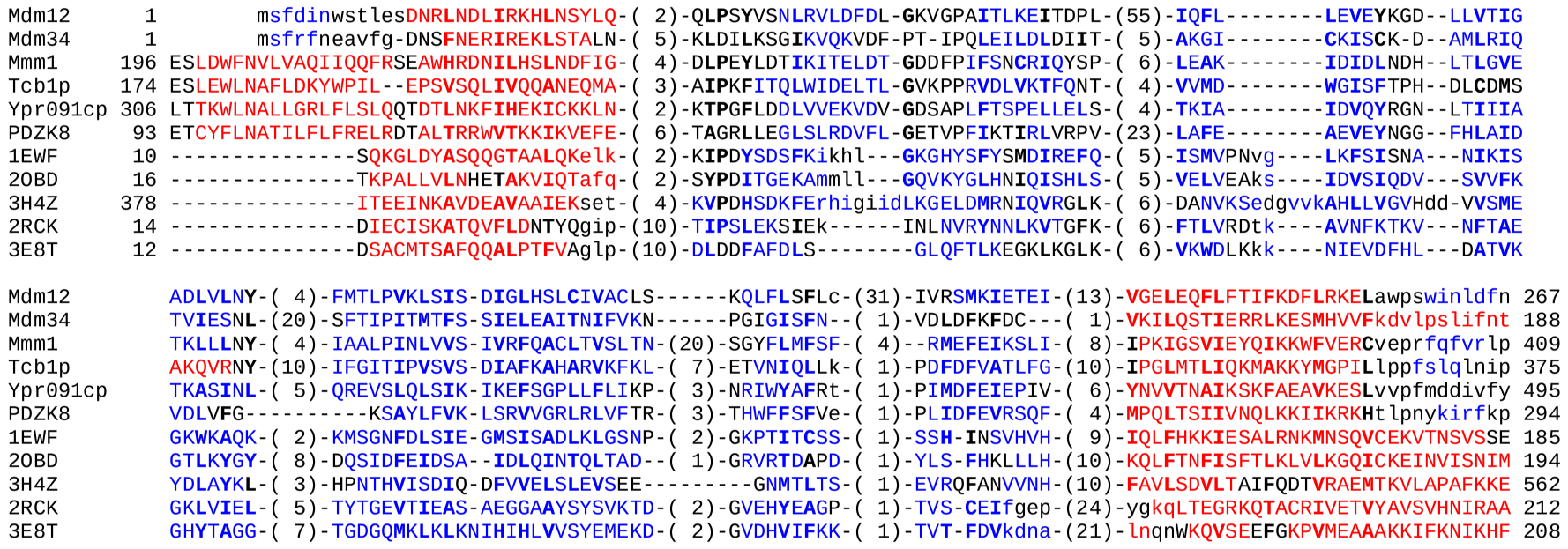


Figure 3.5: Multiple sequence alignment of TULIP domains. An alignment comprising representatives of the SMP domain family and TULIP domains of known structure is shown. Sequences are labeled as in figure 3.2. The α-helices are shown in red and β-strands in blue. Secondary structure predictions for the SMP domains were performed with Ali2D (no version number available; Biegert *et al.*, 2006). The secondary structure of known structure was obtained with UCSF Chimera (version 1.4; Pettersen *et al.*, 2004). Numbers in parentheses represent length of omitted segments. Positions in the alignments that are highly conserved or strongly hydrophobic are shown in boldface. Residues that could not be aligned in structure or sequence are shown in lower case. For details on how the alignment was created, see section 3.2.3.

plication of N-terminal domains of BPI-like proteins in insects. These and the arthropod allergen proteins without tandem duplications are connected to BPI as well as Takeout-like proteins, bridging the two groups (yellow and orange clusters in figure 3.6). With these family interconnections, we were interested to analyze the three families in detail based on the proteins in our cluster map.

3.3.4 *The BPI-like family*

The core group of this family is formed by the closely related BPI, LPSBP, PLTP, CETP, BPI-like 2, and lipid-binding proteins from plants. Although grouped in a fairly tight cluster in figure 3.6, these proteins can be separated into individual clusters with more stringent clustering criteria, showing that the divisions implied by the different names are not evolutionarily arbitrary. The tightest relationship is seen between BPI and LPSBP, which have a common outgroup in

Cluster	gi		
BPI-like family			
BPI/LBP	2653817	471241	27155085
phospholipid transfer protein (PLTP)	14583090	119596192	56681181
cholesteryl ester transfer protein (CETP)	126031487	109128687	1345733
BPI-like 1/LPlunc2	109092363	73992203	109730991
BPI-like 2	281182880	76662810	148689482
BPI-like 3	34395532	151357875	151357875
lipid-binding protein (LBP)	222641797	226507306	218202346
LPLUNC	194224317	281339294	114681506
SPLUNC	27806071	73991580	12845383
Giardia proteins	253746468	159108077	159117809
ESAG5 proteins	261334797	189094729	197090948
phylogenetically diverse	66814342	281201906	196008865
fungal proteins	213410609	261190206	212532739
nematode proteins	268576004	282158126	268537404
Takeout-like family			
Takeout/JHBP	112983172	194277477	145284387
insect proteins with duplicated TULIP	195359219	125985295	242008749
Ixodes (tick) proteins	241600171	241672717	241999240
allergens	14423650	37958151	33772596
allergen outgroup	156544901	238908542	241629381
SMP domain-like family			
Mmm1	259496081	115386312	162312188
Mdm12	259495522	213408148	254570741
Mdm34	259495582	261193022	212538239
PH domain-containing	168032429	270239157	194216759
PDZ domain-containing	73998896	164698472	126273380
synaptotagmin-like	238485966	212526132	293349410
transmembrane 24	122891010	224042537	119587861

Table 3.2: Protein accession codes (gene identifiers; gi) of three representatives of each cluster in figure 3.6.

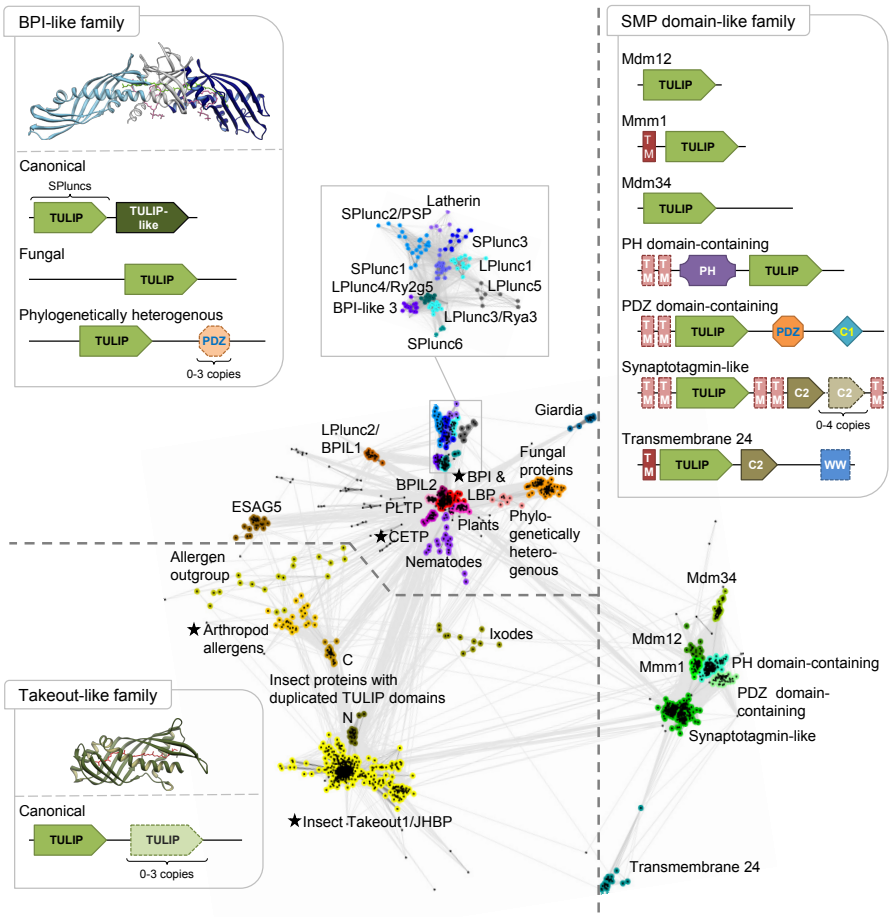


Figure 3.6: Cluster map of the TULIP domain superfamily, computed with CLANS from all-against-all pairwise BLAST p-values (section 3.2.2). Dots represent sequences. Line coloring reflects p-values; the brighter a line, the lower the p-value. Grouped sequences are shown in the same color; sequences that could not be assigned to a group are shown in black. Broken lines divide the cluster map into the BPI, Takeout, and SMP domain-like families and data on structure and domain composition of the groups are shown. Groups without explicit domain composition are canonical; individual sequences might have compositions that differ from the ones shown for their groups. Broken outlines indicate domains present in some but not all proteins of a group. The blow-up shows a clustering of only the PLUNC group. The structure shown as representative of the BPI group is BPI (1EWF); the one shown for the Takeout/JHBP cluster is Takeout 1 (3E8T). Groups of known structure are marked with a star. Accession details for representatives of all clusters are provided in table 3.2.

proteins from fish, suggesting that they resulted from a duplication event at the root of terrestrial vertebrates.

Several satellite groups radiate from this core group. One of these contains the PLUNCs (see blow-up in figure 3.6), which are mainly found in two clusters. One is closer to BPI and contains LPLUNC3/Rya3, LPLUNC4/Ry2g5, SPLUNC6, and BPI-like 3, and one is more divergent and includes LPLUNC1, SPLUNCs 1–3 and latherin. We conclude that the SPLUNCs are polyphyletic, with SPLUNCs 1–3 and latherin originating by deletion from LPLUNC1 and SPLUNC6 from LPLUNC4. LPLUNC5 is clearly separate and approximately equidistant to these two groups. Outside the PLUNC group and making connections of about equal statistical significance to PLUNCs and to BPI lies a group of proteins annotated as LPLUNC2 and/or BPI-like 1, which thus appear to represent a separate evolutionary development from both PLUNCs and BPI.

Three other satellite groups are formed by uncharacterized proteins that are either entirely or largely genus-specific (figure 3.6): the aforementioned ESAG5 proteins from Trypanosomes, a group of proteins from *Giardia*, and a group of nematode proteins, mainly from *Caenorhabditis elegans*.

All proteins of the BPI-like family mentioned so far, with the exception of SPLUNCs, are formed by a tandem of N-terminal TULIP and C-terminal TULIP-like domains, and lack additional domains. The architecture is however different in the last important satellite group, in which proteins are characterized by their large size (about twice the size of other proteins in the BPI-like family) and by the fact that they only contain the TULIP domain, typically towards their C-terminus. This group consists of an intermediate cluster of phylogenetically heterogeneous proteins from slime moulds, diatoms, and amoebae and a more divergent cluster of proteins almost exclusively from fungi, which itself separates into two paralogous subclusters at higher clustering stringency. Three of the proteins in the intermediate cluster contain three tandem PDZ domains C-terminally to the TULIP domain; otherwise, the domains of these proteins could not be annotated with current databases.

A number of additional proteins and protein clusters radiate from the core group, which we have not labeled at this time. They originate mainly from deeply branching eukaryotes (amoebae, ciliates, slime moulds, choanoflagellates, kinetoplastids, unicellular green algae), but also from nematodes. Most show the tandem of TULIP and TULIP-like domains typical for BPI-like proteins, but several are very large (about a thousand residues), contain only the TULIP domain in single or double copy at their N-terminus have an extended transmembrane region with nine predicted transmembrane helices at their C-terminus.

3.3.5 *The Takeout-like family*

This family mainly consists of two groups of sequences. The larger of these two is further removed from the BPI-like family and contains insect proteins, many of which are annotated as Takeout and/or juvenile hormone-binding protein (JHBP). The name-giving protein of this family, *Drosophila* Takeout, connects circadian rhythms with feeding behavior (Sarov-Blat *et al.*, 2000) and also affects male courtship behavior (Dauwalder *et al.*, 2002). The structure of its ortholog from the moth *Epiphyas postvittana* in complex with ubiquinone-8 shows a fold very similar to that of the N-terminal half of BPI-like proteins, with the ligand bound in the same place within the central tubular cavity (3E8T; Hamiaux *et al.*, 2009). Few other Takeout homologs have been characterized to date, but various findings suggest that many may be involved in chemosensory perception (Fujikawa *et al.*, 2006) or hormone delivery (Gilbert *et al.*, 2000). The best understood of these are the JHBP of Lepidoptera, which bind the terpenoids that control insect life cycle in the haemolymph and deliver them to the target tissues . The crystal structure is known for two of these, one from silk worm in complex with juvenile hormone III (2RQF; Suzuki *et al.*, 2011) and the other from honeycomb moth (2RCK; Kolodziejczyk *et al.*, 2008). The structures are again very similar to the TULIP fold, including the mode in which the hydrophobic ligand is bound. This similarity led the authors of the crystal structures to connect the Takeout and JHBP proteins to the N-terminal domain of the BPI-like family, a connection which we could confirm by sequence comparisons. No similarity outside the topology of the fold can be found to the TULIP-like, C-terminal domain of BPI, and this domain must be considered specific to the BPI-like family at this time.

The second main group within this family is closer to the BPI-like proteins and consists of a diffuse collection of arthropod allergens, one of which (dust mite allergen Der p 7; 3H4Z; Mueller *et al.*, 2010), is also of known structure and unsurprisingly shows the TULIP fold. Der p 7 is one of the major causative agents of dust mite allergy in human (Shen *et al.*, 1995; Lynch *et al.*, 1997). Although its exact function is still unclear, it is known that it evokes strong IgE antibody (Shen *et al.*, 1996) and T-cell responses in patients with mite allergy (Thomas and Hales, 2007). Der p 7 was shown to bind bacterial lipopeptide polymyxin B with weak affinity and has been speculated to promote TH2 immunity through co-stimulation of Toll-like receptor 2 pathways (Mueller *et al.*, 2010).

Peripheral to the arthropod allergens is a group of loosely connected proteins, which are closest to the BPI-like family in the cluster map by virtue of making multiple, statistically highly significant connections to the BPI core group. We propose that they represent modern descendants of intermediate stages in the origin of insect Takeout

proteins from a BPI-like ancestor. Several proteins of this outgroup show an up to four-fold amplification of the TULIP domain. Spurred by this observation, we reinvestigated the core Takeout cluster and found individual instances of proteins with multiple copies of the TULIP domain. However, at this time, it is unclear if these proteins arose by duplication and divergence from a single TULIP domain or by fusion.

One group of Takeout-like proteins with two TULIP domains takes an unusual position in the cluster map: whereas its N-terminal domain belongs to the Takeout group, its C-terminal domain forms part of the allergen group. The simplest evolutionary explanation for the origin of these proteins is that they arose by fusion of one TULIP domain from each group. A second explanation is that the location of sequences in the cluster map (figure 3.6) lays out a path for the origin of the insect-specific Takeout proteins from the ancestral BPI-like family, which is common to all eukaryotes. In this second explanation, (1) the group of arthropod allergens originated from the BPI-like family; (2) subsequently one of its members duplicated the TULIP domain; (3) the N-terminal of the two copies diverged away from the arthropod group and (4) became the founding member of the Takeout group through deletion of the C-terminal domain.

The last cluster we found in the Takeout-like family is genus-specific and contains proteins from Ixodes (tick). It seems reasonable to expect that more species-specific clusters will be identified as genome projects provide better coverage of the arthropods.

3.3.6 *The SMP domain-like family*

The SMP domain family comprises eukaryotic membrane-associated proteins. In contrast to the majority of proteins from the BPI- and Takeout-like families that consist solely of TULIP and TULIP-like domains, proteins with SMP domains often contain additional domains. In the original description of the SMP domain (Lee and Hong, 2006), proteins were assigned to groups based either on the nature of these additional domains or on the cellular localization of the proteins (C2 domain-containing synaptotagmin-like, PH domain-containing HT-008, PDZK8, and mitochondrial proteins). As aforementioned, most of these proteins are poorly studied and the SMP domain itself is functionally uncharacterized.

In addition to the aforementioned SMP domain-containing proteins, we detected two further groups in this family (figure 3.6). The first one contains Mdm34 proteins, which are close homologs of the previously known SMP proteins Mmm1 and Mdm12; all three proteins are found only in fungi and are associated with mitochondria. These proteins, like all other members of the SMP domain-like family, have a single TULIP domain, which in Mdm34 is N-terminal and accompa-

nied by an uncharacterized C-terminal domain, in Mmm1 is C-terminal and preceded by a transmembrane helix, and in Mdm12 forms the entire protein (figure 3.1). These three proteins, together with the mitochondrial outer membrane β-barrel Mdm10, form the endoplasmic reticulum-mitochondria encounter structure (ERMES) complex in yeast which we described earlier (section 3.1). It remains unknown whether the complex is merely a tether that allows other proteins to carry out the phospholipid transfer or if the complex itself acts as the transporter. Based on the membership of the SMP domain-like family in the TULIP superfamily of lipid/hydrophobic ligand-binding domains and the abundance of the SMP domain in ERMES, we proposed that this complex might mediate the transport of phospholipids between the ER and mitochondria.

The second group consists of uncharacterized animal proteins annotated as transmembrane 24 (TM24), which are distant homologs of the other SMP domain-like family members. In addition to the TULIP domain, these proteins contain a C2 as well as a WW domain. In the cluster map, they are distant to the other SMP proteins (figure 3.6) as they make most of their connections via a single intermediate sequence from *Branchiostoma*. It is thus at present unclear whether they will move closer to the SMP core group as more intermediate sequences become available through genome projects or whether they will emerge as the founding members of a fourth family of TULIP domains.

3.4 CONCLUSIONS

We have shown that the SMP domain belongs to the TULIP domain superfamily, a large group of proteins that bind lipids and other hydrophobic ligands within a central, tubular cavity (figure 3.7). In several cases (CETP, PLTP), members of this superfamily are known to exploit this binding activity in order to mediate lipid trafficking. Given the extensive lipid exchange between the ER and the mitochondrial outer membrane and the location of the ERMES complex as a connector between them, it is attractive to consider that this exchange is mediated by the SMP domains of the ERMES subunits. As the ERMES complex does not include a nucleotidase that could energize this process, we propose that it proceeds along an affinity gradient, amounting to facilitated diffusion. Although this could be envisaged as resulting from many short, structurally unspecific contacts between the SMP domains (*kiss-and-run* mechanism), we prefer to consider that the domains assemble into structurally well-defined complexes, which establish a lipophilic, tubular path between the two membranes. Since the stoichiometry of subunits within the ERMES complex is currently unknown, it is however not possible at this time to judge on whether

1:1:1 or some other ratio would most appropriately describe the composition of such complexes.

Our results further suggest that the evolutionary roots of this domain lie in the BPI-like family, as it is the only family that contains proteins from basal eukaryotes in addition to those of animals, plants, and fungi. Presumably, the C-terminal TULIP-like domain now prevalent in most BPI-like proteins arose by duplication and diversification of the TULIP domain in early eukaryotic evolution, but its homology to TULIP domains is too distant to be established at this time via sequence comparison methods. Several members of the BPI-like family subsequently lost the TULIP-like domain by deletion; we briefly discussed this in the context of the polyphyletic origin of SPLUNCs. One of these deletion events presumably lies at the root of SMP domain-containing proteins, whose phylogenetic spectrum suggests an origin after the establishment of the eukaryotic cell structure but prior to the emergence of true multicellularity. In a separate deletion event, the Takeout-like family evolved from a BPI-like precursor at the base of the arthropod lineage. Based on the domain structure of various present-day Takeout-like proteins, we propose that this evolution proceeded by consecutive duplication, diversification, and deletion events.

3.4.1 *Recent advancements*

Recently, the crystal structure of an SMP domain was obtained as part of the structure of human extended synaptotagmin 2 (E-SYT2), which tethers ER and plasma membrane (Schauder *et al.*, 2014). Two E-SYT2 proteins form a homodimer by a head-to-head association of their SMP domains, which structurally resembles the arrangement of TULIP

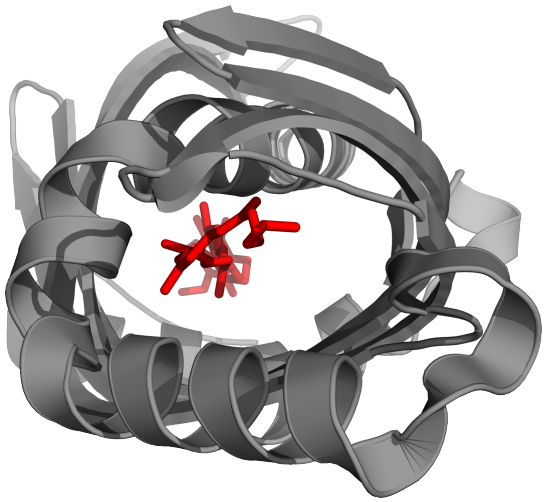


Figure 3.7: View along the ligand-binding tunnel of a Takeout protein (3E8T:5–211). The ligand is shown as red sticks.

and TULIP-like domains in BPI and CETP. The hydrophobic channel was occupied by fatty acid moieties of two lipids in each monomer with their polar heads protruding into the solvent through a seam that covers the whole channel length. The SMP domain of E-SYT2 seems to bind unspecifically to glycerophospholipids, but notably not to cholesterol esters bound by CETP. Three possible modes of lipid transfer are discussed for E-SYT2 based on these novel findings (Schauder *et al.*, 2014). First, a *tunnel* model, in which the SMP domain dimer directly transfers lipids between the ER and plasma membrane by bridging them—this corresponds to our proposed mode of transfer. Second, a *shuttle* model, in which the two membranes are slightly farther apart and E-SYT2 is anchored in both of them with its other domains, allowing the SMP domains to repeatedly traverse this gap. Finally, E-SYT2 need not be the sole effector and additional proteins might contribute lipid specificity to the transfer. More research is necessary to ultimately determine the mode of transfer. Overall, these findings confirm our assignment of SMP domains to the TULIP superfamily experimentally.

4

β-PROPELLER BLADES AS ANCESTRAL PEPTIDES

This chapter includes content from the following publication.
Re-use permissions have been granted by the copyright holder.

Kopec K.O. and Lupas A.N.
β-Propeller Blades as Ancestral Peptides in Protein Evolution.
PLoS ONE, *8*(10): e77074, 2013.

4.1 INTRODUCTION

Previous studies established the homology between proteins of different folds based on the analysis of common fragments (Alva *et al.*, 2008, 2007; Coles *et al.*, 2006; Remmert *et al.*, 2010). These fragments were presumably already found in the last common ancestor of these proteins and were preserved until today even though the proteins themselves underwent fold-changing events. One study found that β-propellers, which adopt toroidal folds comprising 4 to 12 repeats of a 4-stranded β-meander called a *blade*, can be seen for the most part to have arisen by the independent amplification and diversification of one ancestral blade (figure 4.1; Chaudhuri *et al.*, 2008).

The β-strands in each of these blades are named *A* to *D* from the N-terminal innermost strand to the C-terminal outermost one (Chaudhuri *et al.*, 2008). The blades are packed face-to-face and stabilize the fold with hydrophobic interactions (Fülöp and Jones, 1999). In addition, most β-propellers are further stabilized by a *velcro* closure in the first blade, which comprises β-strands from both the N- and C-terminal regions of the domain. Irrespective of the number of blades, β-strands *A*, *B*, and *C* of different blades are usually superimposable with a root-mean-square deviation (RMSD) of below 1Å even though the insertions between these β-strands vary (Chaudhuri *et al.*, 2008).

Proteins of the β-propeller fold are ubiquitous in nature and widely used as structural scaffolds for ligand binding and enzymatic activity (Fülöp and Jones, 1999). To this end, the central funnel-shaped tunnel but also the top, bottom, and lateral regions of different β-propellers can bind ligands (Chen *et al.*, 2011). An interesting example is found in the only solved structure of a ten-bladed β-propeller, Sortilin, where a small peptide forms an additional β-strand that joins the innermost strand of one blade thereby becoming the new edge strand of the β-sheet (Quistgaard *et al.*, 2009).

Given their versatility in forming β-propellers with different blade numbers, it seemed possible that blades may represent ancient peptides that also gave rise to other folds. In this work we therefore extended on previous efforts by including structural information in the detection of β-propeller homologs. We used the aforementioned method of analyzing sequence similarity as a function of structural similarity to distinguish homology from cases of structure-induced sequence similarity. Here, we show the results of these analyses and report on four potential homologous of β-propeller blades.

4.2 MATERIALS AND METHODS

4.2.1 *SCOPβ+*

We created the SCOPβ+ dataset by extending the all-β class of SCOP70, which we chose as a suitable background for distinguishing β-propeller homologs from analogs with similar secondary structure composition (figure 4.2). The 70% sequence identity clustering was obtained from the SCOP version 1.75-based ASTRAL resource (Chandonia *et al.*, 2004). The extension step was necessary to include potential

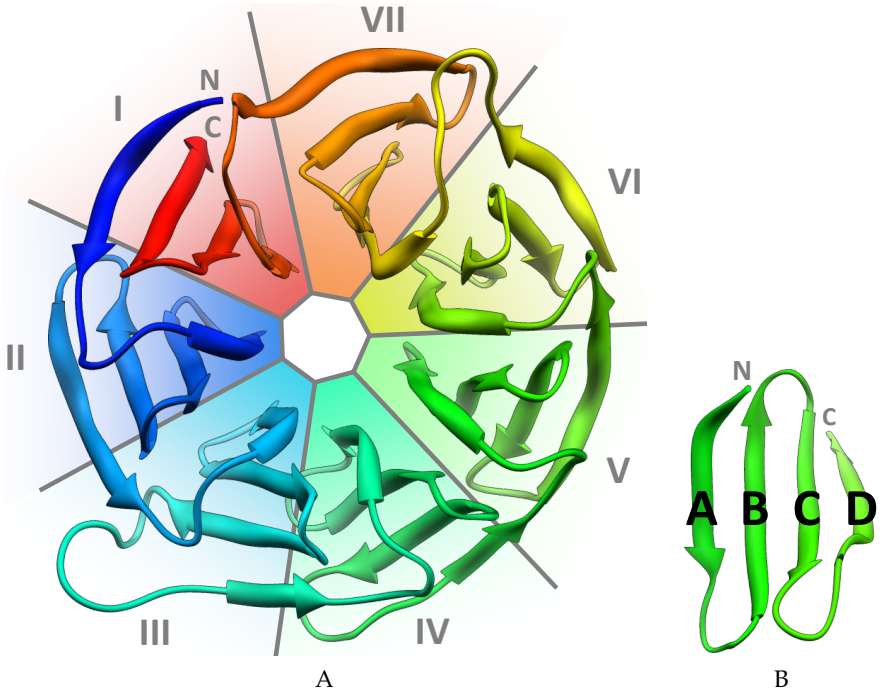


Figure 4.1: Ribbon displays of a β-propeller and a single blade, both with indicated N- and C-termini. A) A 7-bladed β-propeller colored blue to red from N- to C-terminus(2TRC_B:45–340). Blade numbers are indicated. B) A blade is shown with its four β-strands labeled *A*–*D* from N- to C-terminus (2TRC_B:187–223).

β-propeller homologs that are not part of the all-β class of SCOP and to include structures not classified in SCOP.

To establish a scaffold for our extension, we first used the MPI Bioinformatic Toolkit (Biegert *et al.*, 2006) to search the PDB70 database (as available on the 5th of April 2012) using HHpred (HH-suite version 2.0, default parameters; Söding *et al.*, 2005; Söding, 2005), a sensitive remote homology detection method based on the pairwise comparison of profile HMMs. Using a diverse selection of β-propellers from SCOP as query, we recurrently found matches to 4- to 8-bladed β-propellers (folds b.66–b.70), type II β-prisms (BP2; fold b.78), and WW domains (superfamily b.72.1), which we considered the scaffold for our further analysis. The actual extension step started by including all proteins of the all-β class of SCOP70 and extending it by systematically searching PDB70 with all proteins of the aforementioned scaffolding groups. These searches were conducted using the global-alignment mode of HHsearch (otherwise default parameters), the search procedure of HHpred, and matches below 40% probability were discarded. The similarity of some queries led to overlapping matches to the same

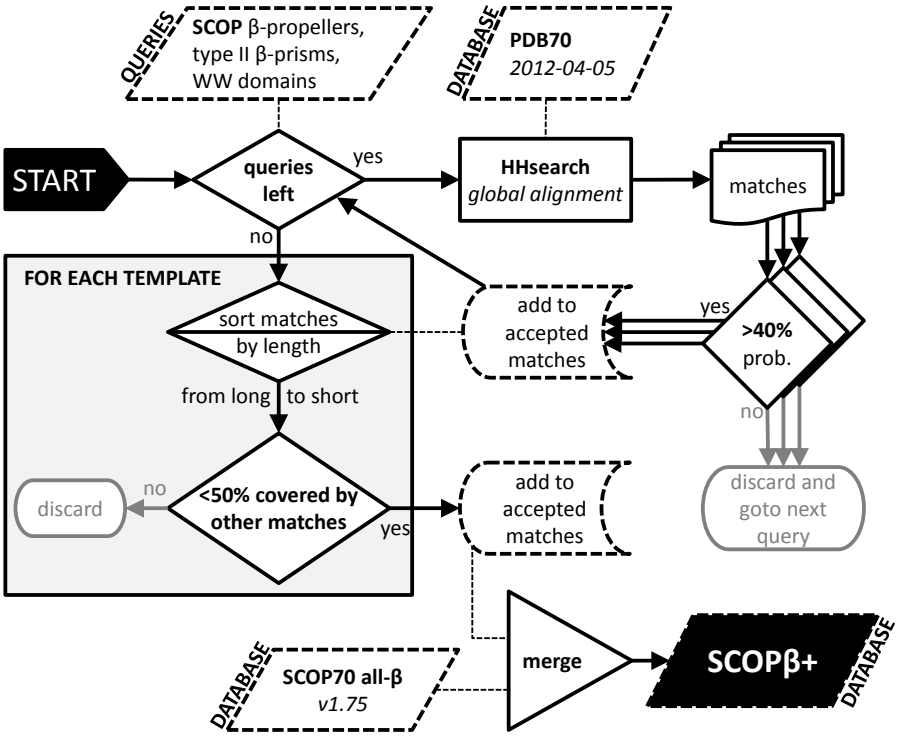


Figure 4.2: First the HHsearch matches of each query against a PDB70 database are filtered and added to a global match list for each template protein. SCOPβ+ is created as the combination of SCOP70 all-β and pseudo-deduplicated template matches, where longer matches are preferred. The start and end points are highlighted. Rectangles are processes, diamonds are decisions, strike-through diamonds are sorting, triangles are merging. Databases and collections of matches have dashed outlines.

template protein. We therefore considered all matches to one template in order of decreasing length and kept only those with more than 50% of their residues not already covered by previously accepted matches. In total, the SCOPβ+ dataset comprises 3223 entries.

A multiple sequence alignment was computed for each entry of SCOPβ+ with the buildali.pl script (a modified PSI-BLAST procedure) and hhmake was used to convert the alignments to profile HMMs—both programs are part of HHpred version 1.6 and were used with default parameters. The profile HMMs of all entries were kept as query profile HMMs and additionally a single database profile HMM was created by merging all of them.

4.2.2 *Cluster map*

We searched the SCOPβ+ database (section 4.2.1) with each query profile HMM using HHsearch in global-alignment mode to obtain an all-vs-all matrix of similarity p-values. These p-values were extracted from the result files and converted to a CLANS (section 2.3) input file using the *bio.io.hhpred* and *bio.io.clans* modules of CSB (version 1.0.0), respectively (Kalev *et al.*, 2012). The cluster map was computed from the input file using the force-directed layouting method implemented in CLANS (no version number available, attract and repulse value 10) at a p-value threshold of $1e^{-5}$ until equilibrium was reached (otherwise default parameters).

4.2.2.1 *Spurious connections*

We found false positive connections in the cluster map and removed them after manual verification (dashed boxes in figure 4.3). A representative example stems from the SCOPβ+ extension search with the N-terminal 7-bladed β-propeller in nitrous oxide reductase (SCOP d1fwxa2) as query. This search resulted in matches to a template protein (3HRP) comprising two domains: a 6-bladed β-propeller and an immunoglobulin-like E set domain. Due to a misaligned match, both template domains were covered and instead of the expected β-propeller domain almost the complete protein was included in SCOPβ+. In the cluster map, this protein was located amidst β-propeller proteins due to its β-propeller domain, but is also—and spuriously so—connected to the immunoglobulin domains.

4.2.3 *Structure-aided homology validation dataset*

To apply the structure-aided homology validation method presented in section 2.5, we first assembled a dataset. For all pairwise structural comparisons we used version 2012/05/07 of TM-align.

First, we created a template dataset consisting of all single-chain SCOP70 entries as well as the β-pinwheels and the inositol-requiring enzyme 1 luminal domain (IRE1-LD) proteins. We created profile HMMs for all 13654 proteins in the dataset as described in the section on cluster map creation (section 4.2.2).

Next, we chose proteins for a *background* dataset, which contains the SCOP all-β class structures that were neither β-propellers nor considered potential homologs of them, i. e. we excluded β-pinwheels, type II β-prisms (BP2s), and WW domains. We used this dataset to evaluate which correlation levels are to be expected for structurally similar yet analogous proteins.

Finally, we assembled a query dataset of 583 blade-like structures from all β-propellers (SCOP folds b.66–b.70), BP2s (b.78), and WW domains (superfamily b.72.1) of SCOP70, β-pinwheel fragments, and IRE1-LD fragments. This dataset contains blades and similar β-meanders that we extracted by manual inspection of the structures.

WW domains were restricted to four residues before the first and three residues after the second conserved tryptophan, similar to their Pfam definition (PF00397; release 26.0; Finn *et al.*, 2014).

The sequences of β-pinwheels are not continuous when considering β-strands *A*–*D* of one blade in structural order. This makes it impossible for TM-align to reasonably align β-pinwheel and β-propeller blades. Thus, we *rewired* the main chain of all β-pinwheel blades by inserting the residues of β-strands *B* and *C* (the putative β-hairpin invasion) in between β-strands *A* and *D* of their blade. We mapped the positions of the reordered residues to the standard β-pinwheels and computed sequence scores using their profile HMMs; this avoided potential problems with generating profile HMMs from artificial sequences.

Both IRE1-LD structures (2BE1 and 2HZ6) contain five potential homologs of blades, however two of them are not in a β-meander conformation but in a long, extended β-hairpin. We excluded the two elongated instances as the structural alignment score for them would not be meaningful and added the remaining three blade-like fragments to the dataset.

As the full-length proteins of all fragments in the query dataset are in the template dataset, we mapped the query fragments onto them, which allowed us to use the template profile HMMs for sequence score computations.

4.3 RESULTS

4.3.1 *β-Propeller homologs*

To detect homologs of β-propellers, we clustered the SCOPβ+ dataset based on pairwise sequence similarities (section 4.2.2). Almost all β-

propellers clustered together, as already observed in a previous analysis of the evolution of β-propellers (figure 4.3; Chaudhuri *et al.*, 2008). For a detailed inspection, we concentrated on proteins with direct or transitive connections to β-propellers at a p-value cutoff of $1e^{-5}$ and omitted all others (figure 4.4A). To annotate groups within this map, we reclustered it at a more stringent cutoff ($1e^{-15}$), which clearly resolved many groups and allowed us to annotate them by manual inspection. The annotations were transferred to the initial cluster map where the groups remained well defined and resolved, also at the less stringent cutoff used for this map (figure 4.4B).

The cluster map depicts the high degree of interconnectedness between different groups of β-propellers (figure 4.4). The biggest cluster acts as a hub for the connections to the outer clusters and is formed by 5- to 8-bladed propellers of known groups: WD40, KELCH, YWTD, YVTN, NHL, PQQ, Clathrin, and PD40 (PF07676) (Chaudhuri *et al.*, 2008; Ghosh *et al.*, 1995; ter Haar *et al.*, 1998). The proximity of these different groups in the cluster map indicates close homology, yet the different groups form distinguishable subclusters.

Adjacent to the hub, three β-propeller clusters are formed by the 4-bladed Hemopexin-like domain family (SCOP identifier b.66.1.1), the RCC1/BLIP-II superfamily (b.69.5), and the loosely connected 7-bladed Sema domain superfamily (b.69.12). Also directly connected to the hub is a large cluster formed by the Asp-Box β-propellers, which are mostly 6- and 7-bladed but also contain the only known 10-bladed β-propeller Sortilin (Quistgaard *et al.*, 2009). The Asp-Box β-propellers are further tightly connected to proteins of the 5-bladed glycoside hydrolase family 43 and more loosely to two 6-bladed Enterobacteria phage K1F β-propellers and to the Integrin cluster.

Interestingly, we found four groups of proteins in the cluster map that are not β-propellers, yet are connected to them: inositol-requiring enzyme 1 luminal domains (IRE1-LD), type II β-prisms (BP2), β-pinwheels, and WW domains. These groups vary in the strength of their connections to β-propellers, from the loosely connected WW domains outgroup to the highly connected IRE1-LD.

In the following sections, we report on our investigations of each of the four folds with respect to an origin from an ancestral blade.

4.3.2 *Inositol-requiring enzyme 1 luminal domains*

The inositol-requiring enzyme 1 luminal domain (IRE1-LD; 2BE1 and 2HZ6) is located within the main β-propeller cluster. This domain detects unfolded proteins in the endoplasmic reticulum (ER) as part of the unfolded protein response and was predicted to adopt a β-propeller fold due to the detection of four blade-like repeats resembling those of an 8-bladed β-propeller (Ron and Walter, 2007; Ponting, 2000). However, both IRE1-LD structures were found to share a unique

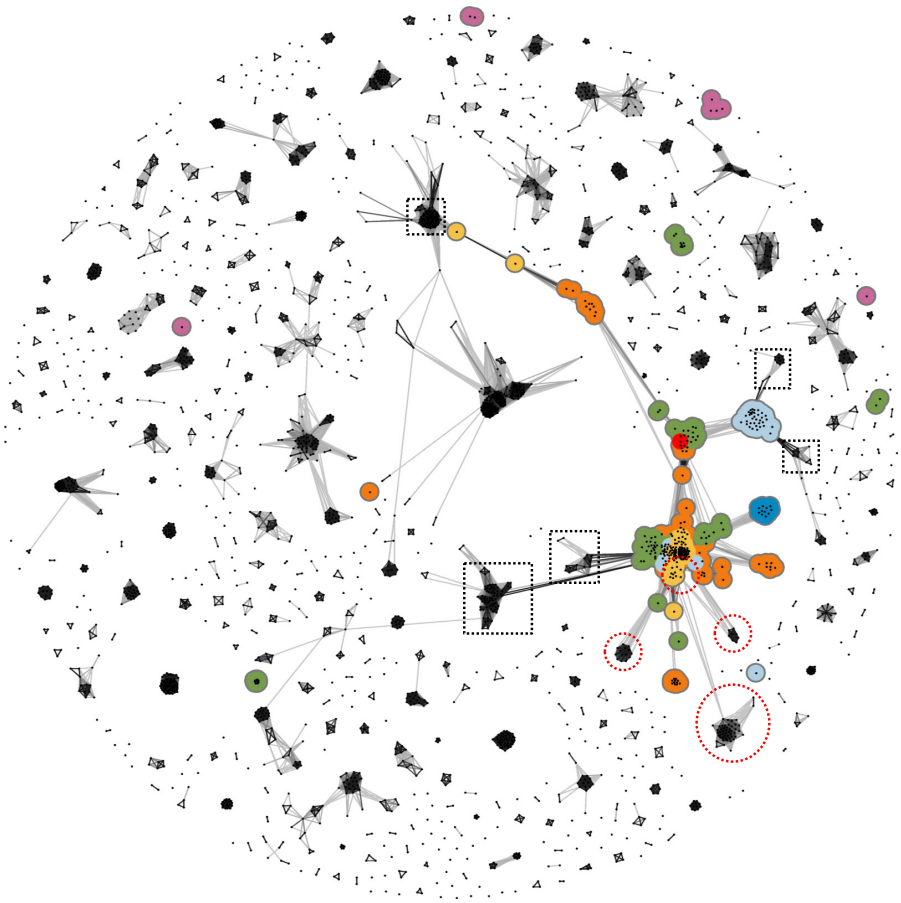


Figure 4.3: Cluster map of SCOPβ+ (Sections 4.2.1 and 4.2.2). β-propellers are colored by the number of blades (4 = blue, 5 = light blue, 6 = green, 7 = orange, 8 = yellow, 10 = red). Most β-propellers are part of one connected cluster network and they are disconnected from most other clusters. A small number of β-propellers, primarily of viral origin, remain unconnected in sequence space, as discussed previously (Chaudhuri *et al.*, 2008). Clusters in dashed boxes were omitted in the detailed analysis after manual inspection (section 4.2.2.1). The purple groups are different superfamilies of the BP1 fold (b.77), unrelated to the BP2 fold discussed here (see figure 4.4). The four clusters discussed in the upcoming sections are in red circles.

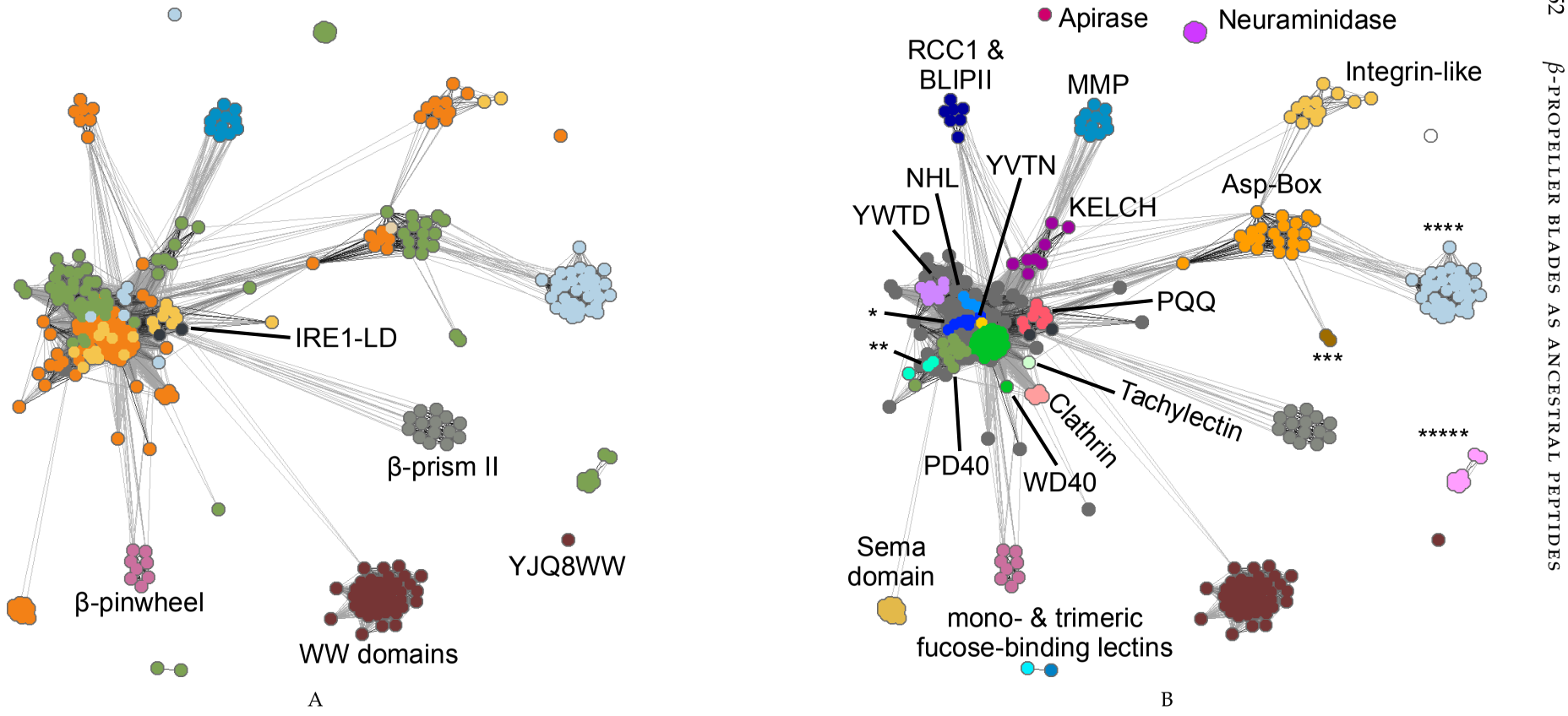


Figure 4.4: Cluster map of β-propellers and their potential homologs (section 4.2.2). Dots represent proteins, connections are similarities where darker means more similar. A) Potential β-propeller homologs are labeled and structural groups are colored as in figure 4.3. The highly divergent WW domain *YJQ8WW* is annotated as well. B) The same cluster map as in A with β-propellers colored and labeled according to motifs and families; names too long to include in the figure are: * = Nitrous oxide reductase N-terminal domain, ** = prolyl oligopeptidase N-terminal domain, *** = Bacteriophage K1F endo-alpha-sialidase, and **** = glycoside hydrolase family 43, ***** = Hemagglutinin-Neuraminidase.

fold that consists of a flat anti-parallel β-sheet, formed by β-strands from two monomers as part of their homodimer interface, and α-helices on one side of the β-sheet that form a groove (Credle *et al.*, 2005; Zhou *et al.*, 2006). Further, the fold has two lobes that are described as a distorted β-barrel and a partial β-propeller for the yeast structure, and as two β-barrels for the human one (Credle *et al.*, 2005; Zhou *et al.*, 2006).

Due to the striking proximity of IRE1-LD to β-propellers in the cluster map, we investigated this relationship in detail. We ran confirmatory HHpred searches with both IRE1-LD proteins as query against the full PDB70 database. The resulting matches were almost exclusively to β-propellers (yeast IRE1-LD: 252 of 258 matches to β-propellers; human IRE1-LD: 332 of 335) and all other matches had low probabilities. Except for a single low-scoring match, the RBB1NT domain of human retinoblastoma-binding protein 1 (2YRV) at 24% probability, all non-β-propeller matches were to WW domains and type II β-prisms (BP2s), both proteins described later in this work. Reverse searches with the top-ranked β-propeller matches confirmed the connection to IRE1-LD.

Next, we were interested in whether state-of-the-art repeat detection methods could automatically detect the four blade-like repeats previously found with a semi-automated procedure (Ponting, 2000). We ran the sensitive repeat detection tool HHrepID (Biegert and Söding, 2008) with the two IRE1-LD sequences as query and both runs detected five repeats. The previously described repeats were the first, third, fourth, and fifth repeat in HHrepID, whereas the second repeat was newly detected. While the first, second, third, and fifth repeat had high probabilities (80%–92%), the probability of the fourth repeat varied between yeast and human IRE1-LD (37% and 89%). However, the sequence segment of this repeat was the same as previously reported and it aligned well to the other repeats.

Mapping the repeats onto the structure revealed that repeats 1, 2, and 5 are three-stranded β-sheets (figure 4.5). In contrast, repeat 3 contains a long central β-strand and two shorter β-strands that form N- and C-terminal $\beta\beta$-hairpins with the central one. Repeat 4 comprises two long β-strands that form an elongated β-hairpin. Repeats 1 and 5 constitute the aforementioned partial β-propeller lobe, whereas repeat 2 is part of the putative distorted β-barrel lobe (Credle *et al.*, 2005). The elongated repeats 3 and 4 are part of the large β-sheet at the homodimeric interface.

To investigate the structural similarity between IRE1-LD repeats and β-propeller blades, we chose the yeast protein as a representative, as a β-strand of repeat 3 in the human structure is not solved. We superimposed repeats 5 and 1 onto two consecutive blades of the 8-bladed BamB β-propeller (3Q7M), which was the top match in the aforementioned HHpred run. Interestingly, this also superimposed the C-terminal β-hairpin of repeat 3 to the third consecutive blade of the

β-propeller, i. e. repeats 5, 1, and 3 are alignable to three consecutive blades. The superimposition aligns the three repeats to the outer blade β-strands, which is peculiar given that strand *D* is known to be the structurally least conserved one in β-propellers (Chaudhuri *et al.*, 2008). The newly detected repeat 2 is slightly more distorted than repeats 1 and 5 and therefore did not align as well to β-propeller blades. In a superimposition of repeat 2 and one BamB blade, repeat 3 again comes close to the subsequent β-propeller blade, albeit not as well as when repeats 5 and 1 are used to set the superimposition.

The aforementioned BamB, along with many other top matches of the IRE1-LD HHpred searches, belongs to the PQQ family of β-propellers. These proteins contain an 11 residue motif on β-strands *C* and *D* of each blade, which ends with a tryptophan at position 11 (figure 4.6; Ghosh *et al.*, 1995). The motif comprises two key structural components: (1) residues 6 and 7 of one blade are arranged parallel to the indole ring of Trp11 from the previous blade and (2) the main chain carbonyl of residue 4 is hydrogen-bonded to the Trp11 indole NH group within the same blade (Ghosh *et al.*, 1995). We analyzed IRE1-LD with respect to these two features and found that they are mostly conserved in the structural interactions between repeats 5, 1, and 3. In yeast IRE1-LD, repeat 1 interacts with both structural neighbors, whereas in human IRE1-LD only the interaction between repeats 5 and 1 is seen due to missing density. The more distorted repeat 2, as well as the elongated β-hairpin-like repeat 4 do not show these characteristics. As the conserved residues of PQQ β-propellers are located in β-strands *C* and *D*, and play a structural role, it is less surprising that the IRE1-LD repeats align to the outer β-strands and not to the usually well-conserved β-strand *A*.

To further verify these findings, we applied a method that analyzes the correlation between structure and sequence similarity (in the following: sequence-structure correlation; see section 2.5 and figure 4.7). We omitted IRE1-LD repeats 3 and 4, as their elongated β-hairpin-like structures make them unsuitable to compute sensible structural alignments to β-propeller blades. The correlation between structure and sequence similarity scores when comparing the IRE1-LD repeats to the background set (section 4.2.3) was 0.11 (median TM-score and HHalign score: 0.38 and −0.18). As the background set is a subset of the SCOP all-β class, a low non-zero correlation was to be expected due to shared β-strand propensity. In comparisons of IRE1-LD to β-propellers with different blade numbers, 8-bladed β-propellers had the highest correlation value (correlation 0.72, TM-score 0.60, HHalign 0.59). We found that the overall highest correlation was achieved in the comparison to the aforementioned PQQ subset of 8-bladed β-propellers and these comparisons also had remarkably high sequence similarity scores (correlation 0.89, TM-score 0.63, HHalign 1.17).

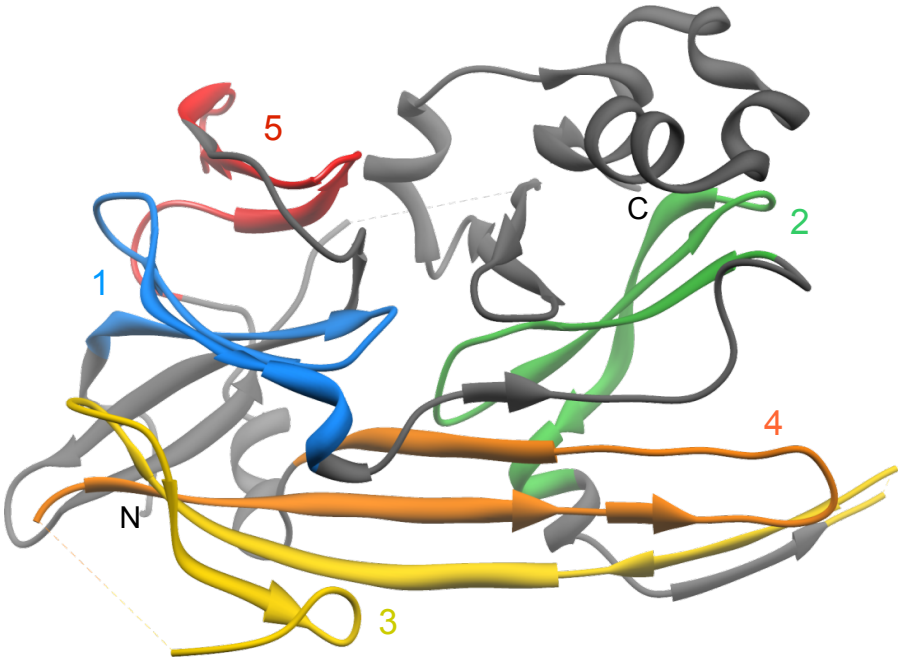


Figure 4.5: Structure of the yeast IRE1-LD monomer with the five repeats detected by HHrepID colored and labeled.

```
                          1 1
                1234567890 1
PQQ motif       AXDXXTGDXX W
                  N   E
                      K
BamB - blade 2  ALNADDGKEI W
BamB - blade 3  ALNTSDGTVA W
BamB - blade 4  ALNEADGAVK W
BamB - blade 5  AVLMEQGQMI W
BamB - blade 6  ALDLRSGQIM W
BamB - blade 7  ALTIDGGVTL W
BamB - blade 8  WINVEDGRFV W

y_IRE1-LD – r1  AVDRRNGHII W
y_IRE1-LD – r2  YFNAHQGLQK L
y_IRE1-LD – r3  TINMLNGEII S
y_IRE1-LD – r4  IHSYDGASYN V
y_IRE1-LD – r5  ASDLDFRIAR W

h_IRE1-LD – r1  AVSKRTGSIK W
h_IRE1-LD – r2  TLGSKNNEGL T
h_IRE1-LD – r3  VIDLLTGEKQ Q
h_IRE1-LD – r4  MYDTKTRELR W
h_IRE1-LD – r5  TVDSESGDVL W
```

Figure 4.6: Sequence alignment of the 11-residue PQQ motif in blades 2–8 of BamB (3Q7M; blade 1 is a velcro blade and was omitted for not being continuous in this region) with the corresponding regions of yeast (y) and human (h) IRE1-LD repeats 1–5. At the top is shown the consensus PQQ motif (Ghosh *et al.*, 1995). Conserved motif positions have a gray background and residues adhering to the consensus are highlighted.

Even though IRE1-LD adopts a fold that is globally different from a β-propeller, our analysis indicates that IRE1-LD is closely related to PQQ β-propellers. The antecedent blades are still detectable as repeats even though they only have three β-strands remaining or have changed their conformation. The complexity of the IRE1-LD fold and the five PQQ-like repeats make it unlikely that this fold has arisen by amplification from a single blade. Instead, it is conceivable that a PQQ β-propeller underwent a massive fold change, which was retained due to its emergent usefulness in ER stress sensing.

4.3.3 *Type II β-prisms*

The second group of potential β-propeller homologs in our cluster map are type II β-prisms (BP2; SCOP fold b.78). Proteins with this fold form a superfamily of phylogenetically widespread lectins, referred to as *Galanthus nivalis* agglutinin-related lectins (GNA-related lectins) after the first structure of this fold (Hester *et al.*, 1995). The BP2 fold comprises three four-stranded β-meanders that are arranged around and orthogonal to a central pseudo-symmetry axis and are curved towards the center (figure 4.8). Similar to β-propellers, which circularly permute between one and three β-strands of a terminal blade in order to hydrogen-bond their N- and C-termini and achieve increased stability (velcro closure), BP2 proteins also use velcro closure for their domain organization and dimerization (Hester *et al.*, 1995; Chandra *et al.*, 1999). The sugar-binding motif is located on the outer, concave side of up to three of the β-sheets (Ramachandraiah and Chandra, 2000; Shetty *et al.*, 2012). Even though sugar binding is their most discussed function, GNA-related lectins also possess 1) anti-tumor, antifungal, and anti-viral activity, 2) bind the HIV surface glycoprotein GP120, and 3) can be taste modifying (De Mejía and Prisecaru, 2005; Li and Romeis, 2009; Hoorelbeke *et al.*, 2011; Kurimoto *et al.*, 2007).

It is important to discriminate BP2 from the type I β-prism (BP1; SCOP fold b.77), which resembles BP2 structurally but has β-strands running parallel to the pseudo-symmetry axis. BP1 proteins also bind carbohydrates with up to three binding sites, and a common origin of BP1 and BP2 has been discussed without clear conclusion (Sharma *et al.*, 2008). The large distance between BP1 and BP2 proteins in our cluster map (figure 4.3) indicates that even the most sensitive homology detection methods cannot connect them, thus they should be considered analogs.

The BP2 cluster in our cluster map is an outgroup to the 8-bladed PQQ β-propellers, which are found in the central cluster. A multiple sequence alignment of the three β-sheets of a BP2 (1XD5) and the eight blades of a PQQ β-propeller (BamB, 3Q7M) shows that all three BP2 β-sheets align well with PQQ blades (figure 4.9A). Further, a conserved tryptophan in β-strand 4 of the BP2 β-sheets superimposes,

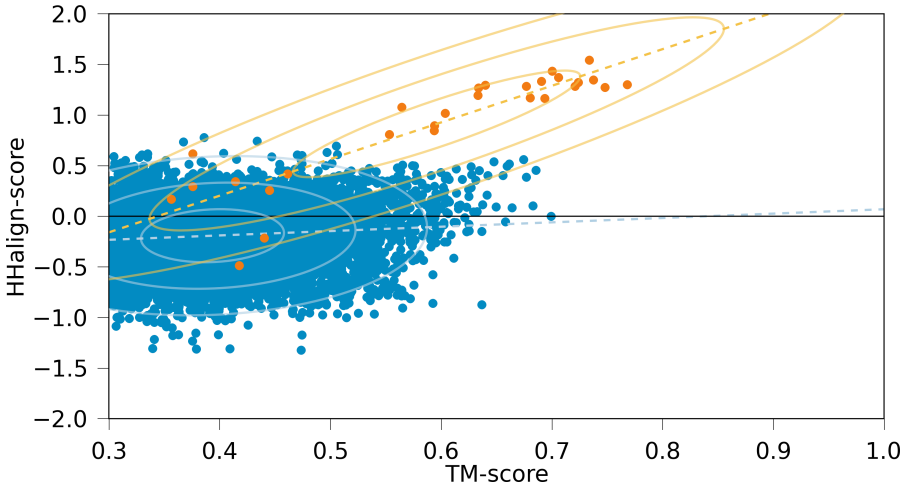


Figure 4.7: Correlation between structure and sequence similarity for the non-propeller fold IRE1-LD and its closest β-propeller superfamily in figure 4.4, PQQ. The panel shows in orange IRE1-LD vs. PQQ. The comparison of the non-propeller fold to a background set of proteins consisting of the SCOP all-β class minus the superfamilies of this work is shown in blue as a reference (see also section 4.2.3). The plots represent the structure (TM-score) and sequence (HHalign) similarity of a pair of compared structures as a dot. The linear regression for each group of comparisons is shown as a dashed line, while the ellipses represent one, two, and three standard deviations around the mean.

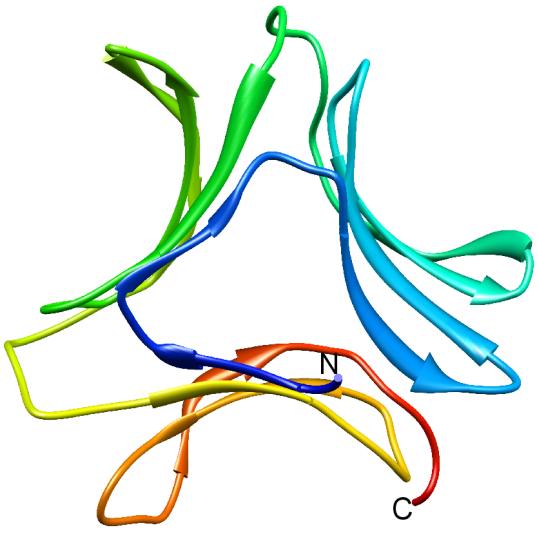


Figure 4.8: Structure of a BP2 (1XD5), colored blue to red from N- to C-terminus.

with slightly different orientation, onto the conserved tryptophan in position 11 of the PQQ specific motif (see IRE1-LD section). The major difference is a two-residue deletion in the BP2 β-sheets, corresponding to positions 5 and 6 in the PQQ motif (figure 4.9B). BP2 may compensate for the missing stabilizing interaction, which residue 6 provides in PQQ motif blades by coordinating the tryptophan side-chain, through the interaction of its three tryptophan residues in the core of the structure (Liu *et al.*, 2005). These differences to the conserved PQQ motif might explain the location of BP2 as an outgroup of PQQ β-propellers.

To verify the presumed homology of BP2 and PQQ β-propellers, we analyzed their sequence-structure correlation (figure 4.10). The similarity scores for structure and sequence comparisons between BP2 and the background set were low and uncorrelated (correlation 0.16, TM-score 0.37, HHalign −0.20). In contrast, the comparisons with 8-bladed β-propellers (correlation 0.45, TM-score 0.53, HHalign 0.48) and their PQQ motif subset (correlation 0.52, TM-score 0.56, HHalign 0.61) showed similarities indicative of a homologous origin of BP2 from PQQ β-propellers. Sequence searches with single BP2 β-meanders as query against PDB70 showed that these are more similar to each other than to any β-propeller blade, suggesting that the BP2 repeats were amplified from a single blade of a PQQ β-propeller.

4.3.4 *β-Pinwheels*

Proteins that adopt the β-pinwheel fold are the third group with connections to β-propellers in our cluster map. They are DNA-binding modules of bacterial type IIA topoisomerases. The first structures with this fold were the C-terminal domains of DNA gyrase A (GyrA, 1SUU) and of the topoisomerase IV ParC subunit (1WP5; Schoeffler and Berger, 2008). DNA gyrase is capable of introducing negative supercoils into DNA, however this function is lost upon removal of either its complete C-terminal domain or of a conserved motif therein, the GyrA box (Kramlinger and Hiasa, 2006; Kampranis and Maxwell, 1996). In contrast, topoisomerase IV, which antagonizes DNA gyrase by relaxing supercoiling, remains functional without the C-terminal domain but loses specificity for positive supercoiling (Schoeffler and Berger, 2008).

Structurally, β-pinwheels resemble β-propellers, with four-stranded β-sheets circularly arranged around a central pore (figure 4.11). Yet the folds differ due to a β-hairpin invasion between neighboring β-pinwheel blades (figure 4.12; Hsieh *et al.*, 2004). Even though they are, strictly speaking, not β-propellers, SCOP classifies them into the 6-bladed β-propeller fold (b.68), where they constitute their own superfamily called *GyrA/ParC C-terminal domain-like* (b.68.10). Interestingly, β-pinwheel structures exist in different variants: completely

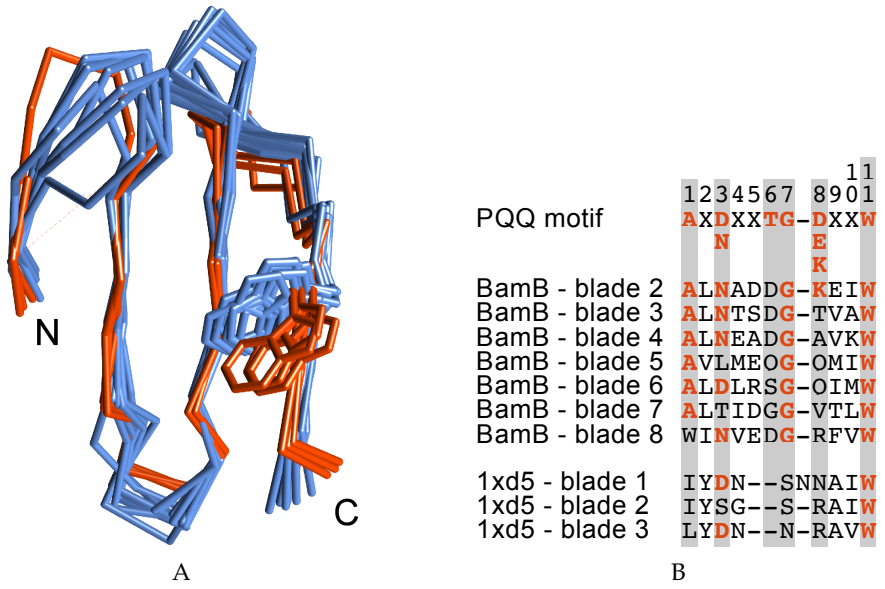


Figure 4.9: Structure and sequence alignments of BP2 β-sheets and PQQ motif β-propeller blades. A) Structural alignment of the three β-meanders of a BP2 (1XD5) and the eight blades of BamB (3Q7M) shown as a main chain trace. The 24 aligned residues result in an average RMSD of 1.28Å. The side-chains of the conserved tryptophan residues in BamB and BP2 are located at the same position but with different orientations. B) Sequence alignment derived from the structural alignment in A. See figure 4.6 for further explanations.

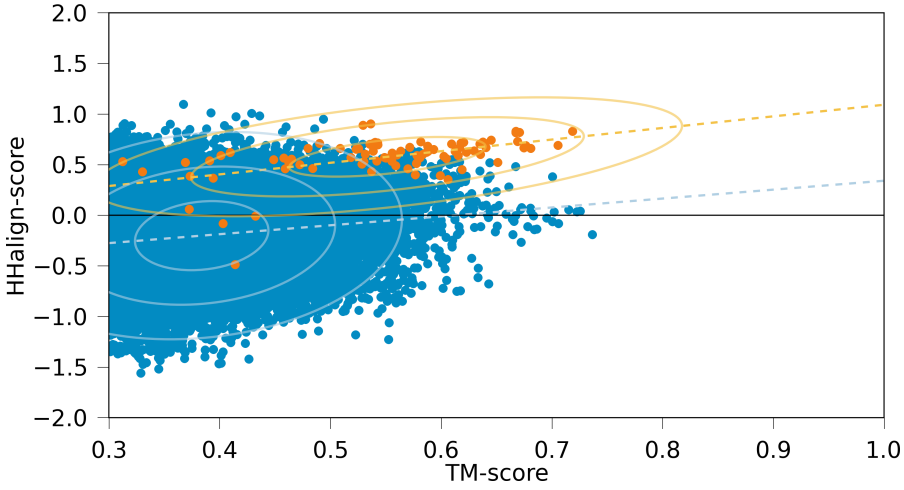


Figure 4.10: Correlation between structure and sequence similarity for the non-propeller fold BP2 and its closest β-propeller superfamily in figure 4.4, PQQ. The panel shows in orange BP2 vs. PQQ. See figure 4.7 for further explanations.

closed circular forms and C-shaped open forms that can be planar or spiral-shaped. It has been suggested that GyrA always has six blades whereas the number in ParC varies from three to eight and it was hypothesized that ParC evolved from GyrA (Corbett *et al.*, 2005).

In DALI searches for structures similar to β-pinwheels, using the C-terminal domains of GyrA and ParC as query, the β-hairpin invasion leads to a clear separation of matches to β-pinwheels (Z-scores > 16) and β-propellers (Z-scores < 5), which are the top matches besides β-pinwheels (Holm and Rosenström, 2010). In these searches, the 6-bladed closed β-pinwheels were most similar to 6-bladed β-propellers, whereas the C-shaped forms with five or six blades had 7-bladed β-propellers as top matches.

In these searches, we found six additional β-pinwheel domains (1ZI0, 1ZVU, 1ZVT, 3L6V, 3NO0, 3UC1) and queried PDB70 with all eight β-pinwheels using HHpred. We pooled the results into a non-redundant list and, after the self matches, 33 and 3 of the following 40 matches were to 7- and 8-bladed β-propellers, respectively, and only 4 low-scoring matches were to proteins of other folds. The majority of the β-propeller matches were to 7-bladed β-propellers with the WD40 motif, which is in agreement with the cluster map, where β-pinwheels almost exclusively connect to WD40 β-propellers. For confirmatory reverse searches, we used the 10 best β-propeller matches. In all cases, the best β-pinwheel match had a probability above 50% and in eight of ten searches above 80%. All reverse searches matched multiple β-pinwheels and the matches were interspersed with matches to various β-propeller groups. An earlier study had proposed RCC1 as the group of β-propellers with the highest similarity to β-pinwheels, but our analysis indicates only a transitive connection between these groups via the proteins of the main β-propeller cluster, a finding consistent with the previously noted lack of key RCC1 residues in gyrase A (Qi *et al.*, 2002; Stevens and Paoli, 2008).

Due to the rather low sequence similarity of β-pinwheels and WD40 β-propellers, which is also evident from their distance in the cluster map, it is not surprising that the WD40 motif-defining tryptophan and aspartate residues are not conserved in β-pinwheels.

To investigate whether the sequence similarity between β-pinwheels and β-propellers could be structure-induced, we again computed sequence-structure correlations (figure 4.13). Due to the β-hairpin invasion, TM-align is unable to align β-pinwheel and β-propeller blades in a reasonable way; therefore we created artificially reordered β-pinwheel blades (section 4.2.3). The correlation of structure and sequence similarity between the reordered β-pinwheels and the background set was 0.12 (TM-score 0.39, HHalign -0.13), which is in line with the results for IRE1-LD and BP2. The correlation of scores between the reordered β-pinwheels and the WD40 β-propellers, which were their best sequence matches, was indistinguishable from the background

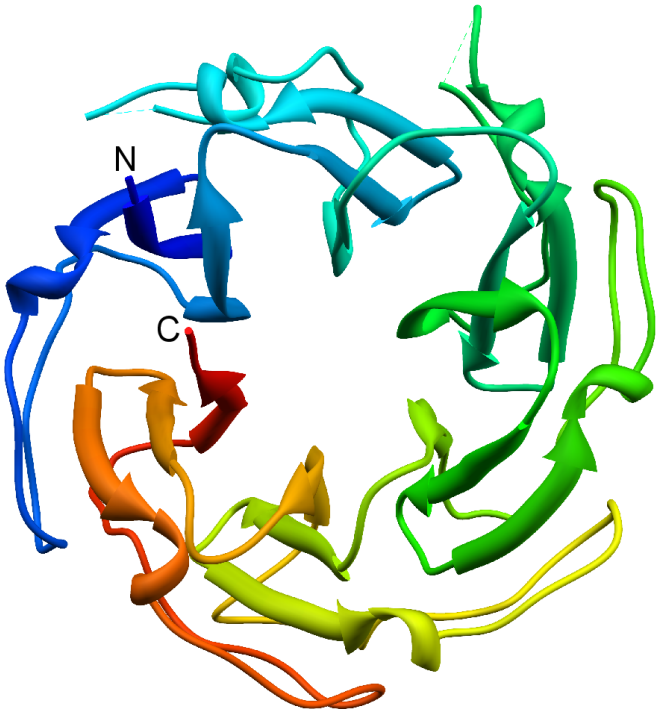
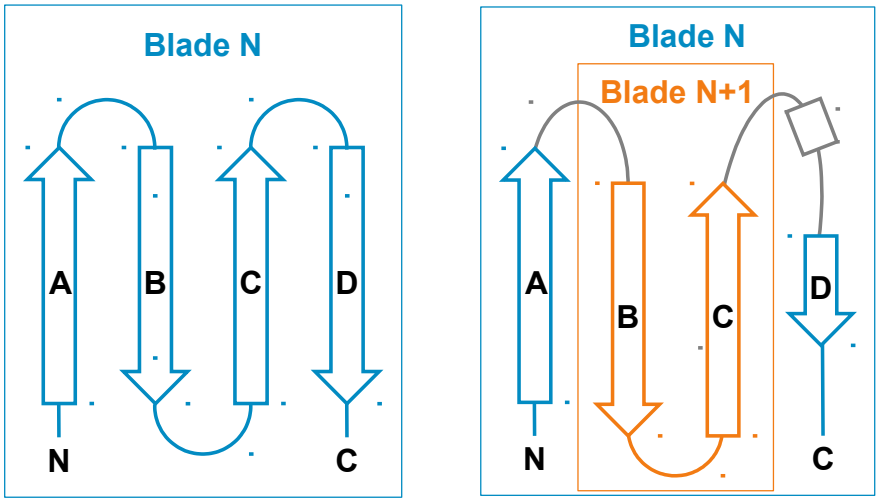


Figure 4.11: Structure of a closed-form β-pinwheel (1SUU).



A β-propeller B β-pinwheel

Figure 4.12: Topology diagrams of four consecutive β-strands in A) β-propellers and B) β-pinwheels. In β-propellers, the four β-strands form a single β-propeller blade. In β-pinwheels, β-strands *B* and *C* are part of the next blade, such that the four consecutive β-strands are part of two blades.

(correlation 0.12, TM-score 0.56, HHalign 0.37). Both scores are higher than for the background set, but there is no significant correlation between them, indicating that the sequence similarity may be structure-induced and thus pointing to a convergent origin of WD40 and β-pinwheels, as previously proposed (Corbett *et al.*, 2004).

The apparent similarity of β-pinwheels to β-propellers in sequence searches may be due to the two folds being formed by repeats of the same length and secondary structure. This is because the statistical significance of comparisons between repetitive proteins increases with the number of repeats that can be matched, even when the repeats individually have little or no detectable similarity. In this case, searches with single reordered β-pinwheel repeats did not show even low-scoring matches to β-propellers. We therefore conclude that this similarity is not indicative of homology.

4.3.5 *WW domains*

The fourth group we found connected to β-propellers in our cluster map is the WW domain superfamily (b.72.1). Members of this superfamily adopt a ~38 residue long fold comprising a curved three-stranded β-meander with two highly conserved tryptophan residues (Bork and Sudol, 1994). The N-terminal of these is located in the first β-strand and projects to the convex side of the β-sheet, whereas the C-terminal is in the third β-strand and has its side-chain on the concave side. Together with a conserved tyrosine in the central β-strand, the latter forms a binding site for proline-rich motifs (figure 4.14; Sudol *et al.*, 2005). WW domains are known to occur in tandems of up to four copies and one reason for this amplification might be to increase binding affinity (Hofmann and Bucher, 1995; Webb *et al.*, 2011). Structurally, a WW domain corresponds to three β-strands of one β-propeller blade.

In our cluster map, WW domains are loosely connected to the main β-propeller hub and HHpred searches with single domains often had β-propellers as low-scoring matches, with similar results for the reverse searches. Since, as mentioned for β-pinwheels, the statistical significance of comparisons between repetitive proteins increases with the number of repeats that can be matched, we decided to compare searches with single domains to searches using several domains in tandem.

Searches of single WW domains (1E0L, 1E0N, 1PIN, 1WR4) with HHpred against PDB70 yielded matches to IRE1-LD and several β-propellers, scattered sparsely among other matches and mostly with probabilities below 40% (but occasionally as high as 70%). Although the second conserved tryptophan was in some cases aligned to the conserved tryptophan of PQQ β-propellers and IRE1-LD, many high-scoring matches did not have conserved residues at this position.

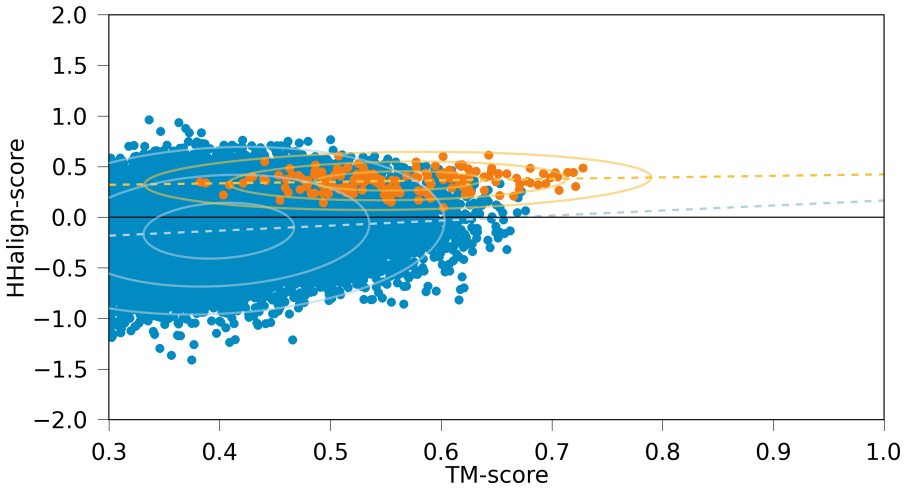


Figure 4.13: Correlation between structure and sequence similarity for the non-propeller β-pinwheel fold and its closest β-propeller superfamily in figure 4.4, WD40. The panel shows in orange β-pinwheels vs. WD40. See figure 4.7 for further explanations.

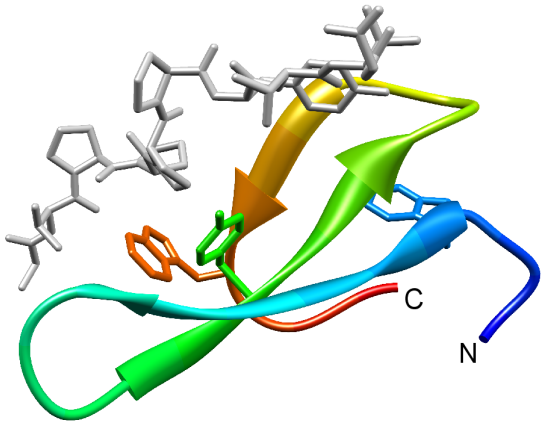


Figure 4.14: Structure of a WW domain (1JMQ_A:13–42) bound to a proline-rich peptide (gray). The side-chains of the conserved tryptophan residues and the binding-site tyrosine are shown.

Searches of double WW domains (1O6W and 2JXW) showed an increase in number and probabilities of matches to IRE1-LD and β-propellers, particularly to the 8-bladed PQQ β-propellers (up to 93%). Here, two consecutive blades frequently aligned without or with only few gaps to the query WW domains and the conserved C-terminal tryptophan residues in each repeat were aligned.

Searches of quadruple WW domains confirmed our previous results (gi | 73919464:363–554, 2072503:300–477, 73921204:193–581). Here again, BamB was among the top β-propeller matches (88% probability) and it covered the four WW domains with four consecutive blades, the conserved PQQ motif tryptophan of all four blades being matched to the second WW domain tryptophan.

To assess the structural similarity of WW domains and PQQ motif blades, we compared a double WW domain (1O6W) to its top-matching β-propeller, the 8-bladed BamB, in structure and sequence (3Q7M; figure 4.15A and 4.15B). The superimposition had an RMSD of 1.9Å over the three β-strands of the WW domain and the alignment was gapless.

As discussed for β-pinwheels, the tandem domains might have elevated scores due to the alignment of multiple consecutive repeats, which in this case might be further enhanced by the repetition of tryptophan at particular sequence intervals. Hence, this finding is not per se indicative of a homologous relationship.

In order to gain more clarity in the issue of homology vs. analogy, we analyzed sequence-structure correlations (figure 4.16). As in the aforementioned cases, the score correlation between WW domains and the background set was low 0.05 (TM-score 0.38, HHalign −0.41). To our surprise, neither of the β-propeller groups found in the HHpred analysis had significant correlations with WW domains (correlation against PQQ β-propellers −0.18, TM-score 0.46, HHalign −0.04). In conjunction with the sequence searches described above, we conclude that the similarity between WW domains and β-propellers is fortuitous and does not reflect common ancestry.

4.4 CONCLUSIONS

In our search for β-propeller homologs with different folds, we detected four candidate groups: IRE1-LD, BP2, β-pinwheels, and WW domains. These were connected to β-propellers at various levels of statistical significance in sequence comparisons. The question of their evolutionary relationship with β-propellers touches on the problem of distinguishing remote homologs from analogs, a problem that has been discussed for many decades (Fitch, 1970; Russell *et al.*, 1997). In this work we have approached this question by complementing detailed, profile HMM-based sequence comparisons with a recently introduced method that evaluates possible homology based on the cor-

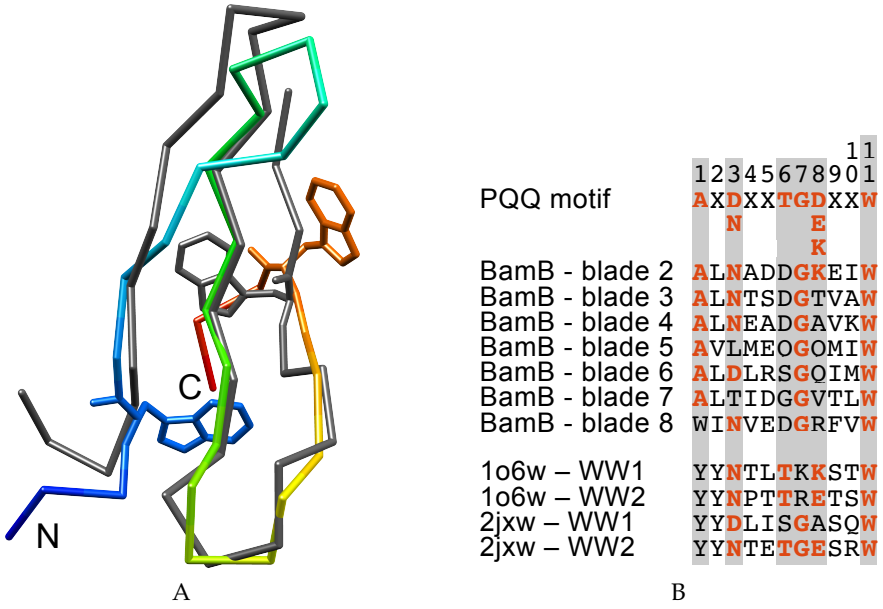


Figure 4.15: Structure and sequence alignments of wwDomains and PQQ motif β-propeller blades. A) Superimposition of the first WW domain of PRP40 (1O6W_A:1–29, rainbow coloring) with β-strands *B–D* of BamB β-propeller blade 2 (3Q7M_A:118–146, dark gray), shown as a main chain trace. The match is gapless and has an RMSD of 1.9Å over 23 residues. B) Sequence alignment of the PQQ motif, BamB blades, and four WW domains. See figure 4.6 for further explanations.

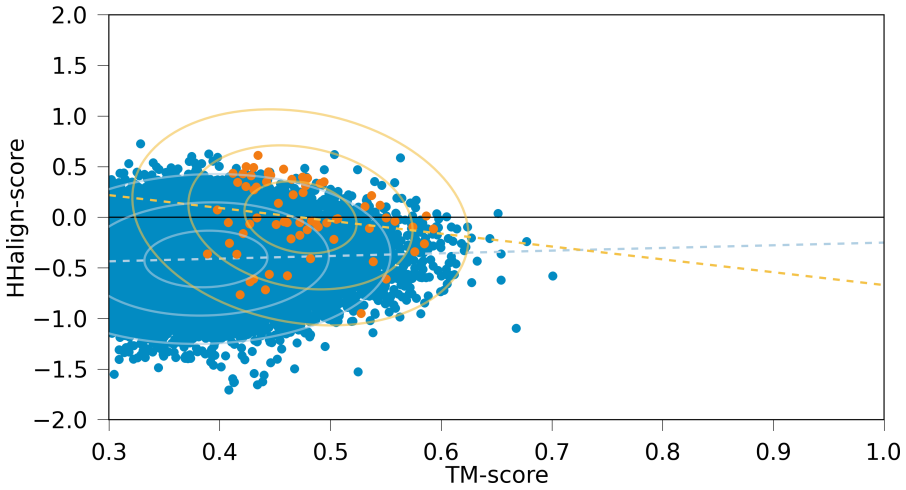


Figure 4.16: Correlation between structure and sequence similarity for the non-propeller WW domain fold and its closest β-propeller superfamily in figure 4.4, PQQ. The panel shows in orange WW domains vs. PQQ. See figure 4.7 for further explanations.

relation between sequence and structure similarity (Remmert *et al.*, 2010). Our results substantiate a homologous relationship between IRE1-LD, BP2, and β-propellers, but indicate that β-pinwheels and WW domains are most likely of analogous origin.

4.4.1 *Evolutionary scenario*

A previous study showed that β-propellers have arisen for the most part by the independent amplification and diversification of one ancestral blade (Chaudhuri *et al.*, 2008). A fundamental question in evaluating the evolutionary relationship of IRE1-LD and BP2 to β-propellers is thus whether they also trace their origin to a single blade. In the case of IRE1-LD, the individual repeats are not more similar to each other than to blades of PQQ motif β-propellers and part of the repeats occur in the same geometry. Overall, the IRE1-LD repeats are so similar to PQQ motif blades that they are found in the same sequence cluster, distinct from clusters formed by other β-propellers (figure 4.4). This suggests that IRE1-LD evolved from a PQQ motif β-propeller by a number of mutations that led to a substantial fold change, rather than by amplification of a single PQQ motif blade. We find that the path taken, however, cannot be reconstructed at this time by concatenation of known fold-changing mechanisms, since no intermediate forms appear to have survived (Grishin, 2001a; Andreeva and Murzin, 2006; Lupas and Koretke, 2008). We note that the part of the IRE1-LD repeats that can still be related to PQQ motif blades by sequence similarity corresponds to blade β-strands *B*–*D*, strand *A* having been replaced in the process of fold change with heterologous segments of the polypeptide chain.

In the case of BP2, conversely, the high self-similarity of its repeating units and their distinctness from the blades of β-propellers indicate a monophyletic origin from an ancestral blade. While it remains unclear whether the BP2 and β-propeller folds arose concomitantly from the same ancestral blade, or whether BP2 emerged subsequently from the amplification of a β-propeller blade that made itself independent of its parent structure, we note that the particular similarity of BP2 to PQQ motif blades suggests the second scenario, with BP2 arising from the blade of a PQQ β-propeller. In this case, again, the part of BP2 repeats that can be related to PQQ motif blades by sequence similarity corresponds to blade β-strands *B*–*D*, strand *A* being formed by an N-terminal extension that completes each repeat consecutively, constraining the structure to an overall triangular shape (figure 4.8). It thus seems possible that the BP2 fold arose by amplification of only the three C-terminal β-strands of a PQQ motif blade and that the N-terminal extension providing the fourth strand to each repeat is of heterologous origin. Experimentally, it may be possible to test the viability of this scenario by attempting to complement triple repeats of

three-stranded β-meanders derived from the C-terminal part of PQQ motif blades with heterologous sequences in a phage display assay. Nevertheless, whether such a process actually led to the emergence of BP2 remains conjectural at this time, as a higher sequence similarity of BP2 repeats to blade β-strands *B*–*D* over other segments of three consecutive β-strands in PQQ β-propellers is not observable.

4.4.2 *Issues in protein classification*

The homologous relationships highlighted here are exemplary for a problem of current protein classification systems. Due to their tree-like structure and their treatment of structural, i. e. analogous, aspects as the prime mean of differentiation, these systems can only represent homologous connections between proteins that share the same fold. Thereby, fold-spanning homology, as in the cases presented here, cannot be captured. To alleviate this issue, the *metafold* was recently proposed as a new classification level, where homologous proteins can be grouped across different folds (Alva *et al.*, 2008). The concept of metafolds can further be applied to bring together proteins that originated from the same ancestral peptide, yet show no global sequence similarity (Alva *et al.*, 2010). Once such a systematic grouping of proteins exists, all analogous criteria could be removed from the classification, which would result in a classification by natural descent.

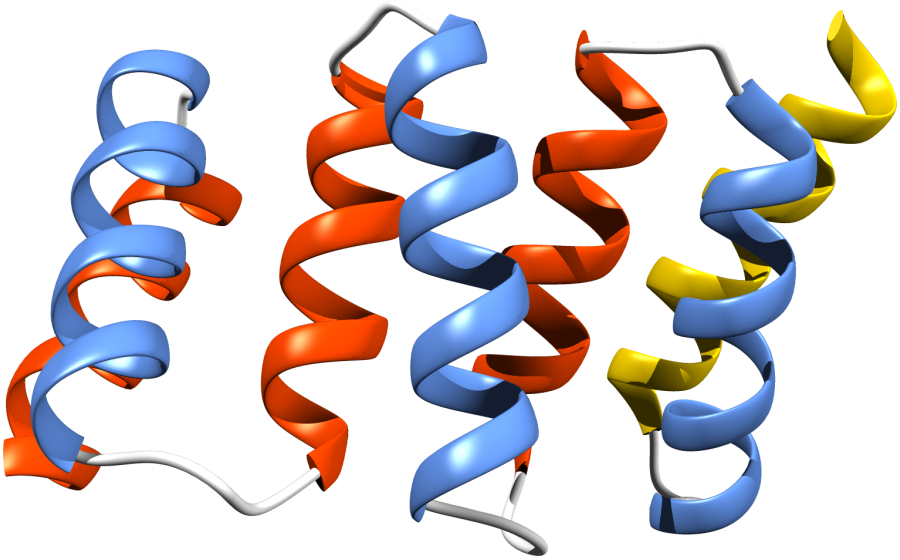
5 EVOLUTION OF TETRATRICO PEPTIDE REPEATS

5.1 INTRODUCTION

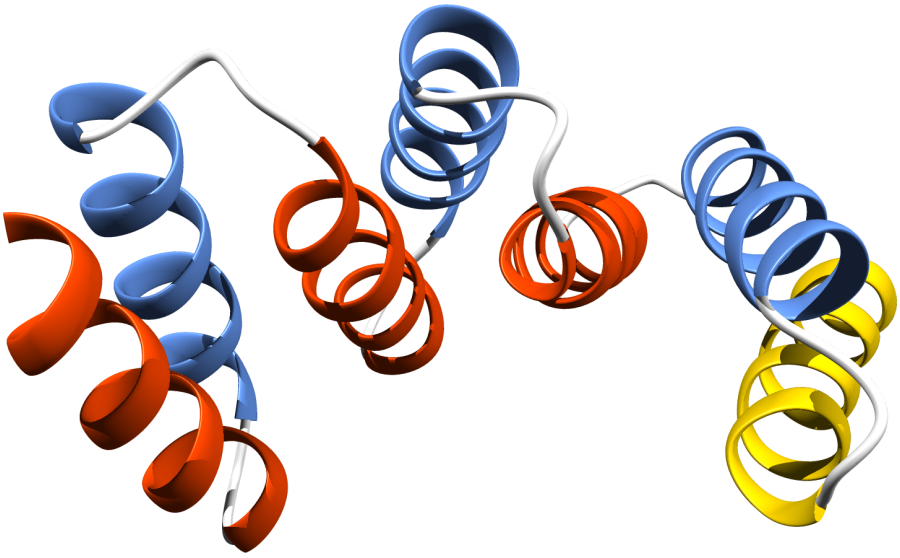
The tetratrico peptide repeat (TPR) is a 34 residue motif encountered in all three domains of life (Sikorski *et al.*, 1990; D'Andrea and Regan, 2003). Each motif instance adopts an $\alpha\alpha$-hairpin conformation with repetitions stacking to form a right-handed and open-ended solenoid referred to as a TPR domain (figure 5.1; Das *et al.*, 1998; Groves and Barford, 1999). In these superhelices, the first α-helix, *A*, constitutes the inner concave surface, whereas both helices, *A* and *B*, contribute to the outer side (D'Andrea and Regan, 2003). The residues on the concave side mediate highly specific protein-protein interactions, making the motif a suitable scaffold for protein engineering (Karpenahalli, 2006; Main *et al.*, 2005, 2003a; D'Andrea and Regan, 2003; Blatch and Lässle, 1999). The usefulness for engineering is also due to the high stability of TPRs, which is increased by *capping* helices that help to avoid solvent exposure of hydrophobic motif residues (figure 5.1; Forrer *et al.*, 2004, 2003; Main *et al.*, 2003b). Following the determination of several, naturally occurring TPR domains, a consensus TPR sequence was derived and the structures of artificial TPR domains with perfectly identical consensus repeats were obtained (figure 5.2; Main *et al.*, 2003b).

Many TPRs are located in TPR domains, however single TPR-like instances are known as well, e. g., in ribosomal protein S20 (RPS20; Karpenahalli, 2006). Ribosomes, and the proteins therein, are of vital importance and thus have been mostly conserved throughout evolution. The single TPR in RPS20 has therefore been discussed as a potential ancestral form of the motif, which was subsequently amplified to repetitive TPR proteins (Karpenahalli, 2006).

Many TPRs can be annotated using homology searches, e. g., by comparisons to the TPR definitions of Pfam or SMART, and with repeat detection methods, however more sensitive methods are available. TPRpred is a custom-tailored approach for TPR detection that derives a profile HMM for each query and compares it to a thoroughly optimized TPR profile (Karpenahalli *et al.*, 2007). Algorithmically, TPRpred is a repeat-aware modification of the HHsearch procedure (section 2.2.5) that allows low-scoring but neighboring repeats to contribute significantly to the overall query score. Unfortunately, the profiles in TPRpred received no post-publication updates and therefore are unlikely to capture the full complement of currently known repeat instances.



A



B

Figure 5.1: Two views of the TPR domain in protein phosphatase 5 (PP5; 1A17_A:28–145). Helices *A* and *B* of each of the three repeats are shown in red and blue, respectively. The terminal solvating helix is yellow. B) shows that *A* helices line the concave surface whereas *B* helices constitute the convex region.

Figure 5.2: A single consensus sequence TPR (tall boxes; *L*1 and *L*2 are link regions) is shown with neighboring *N-cap* (green) and solvating helix (yellow). Highly conserved residues are highlighted including their index in the motif.

In the structural classification of proteins (SCOP), TPRs are found in the TPR-like superfamily (a.118.8) of the $\alpha\alpha$-superhelix fold (a.118). Among the 7 families of this superfamily are several well-known motifs, e. g., TPR itself (a.118.8.1); starch utilization system protein D (SusD)-like domains (a.118.8.6), and HAT/Suf repeats (a.118.8.7). An even more detailed description of TPR-like proteins can be found in Pfam clan TPR (CL0020). Its 117 families give a fine-grained overview of TPR subtypes but also of remotely related TPR-like domains.

Among the resources used for these classifications are studies that analyzed the evolutionary relationship between $\alpha\alpha$-hairpin domains and TPRs (e. g., Zhang and Grishin, 1999). While these exemplary findings are interesting, we wanted to compose a more complete view of the evolution of TPR-like $\alpha\alpha$-hairpins. To this end, we first investigated whether we could find recently amplified TPRs that would hint at a still ongoing process of TPR domain genesis from non-repetitive motif instances. Apart from this focus on TPR domain evolution we further performed in-depth analyzes of other motifs with sequences and structures reminiscent of TPRs.

5.2 MATERIALS AND METHODS

5.2.1 *HHblits*

HMM-HMM-based lightning-fast iterative sequence search (HHblits) is an iterative homology detection program using pairwise profile HMM comparisons (Remmert *et al.*, 2012). Similar to PSI-BLAST, the query alignment (and hence profile HMM) is refined with confidently predicted matches after each round. The drawback of slow pairwise profile comparisons is overcome by a fast initial database reduction step in which templates that are likely to score poorly are discarded. The remaining templates are used in slower but more sensitive pairwise profile HMM comparisons implemented in HHsearch (section 2.2.5) to determine the final results.

The basis of the database reduction step is a set of 219 pre-defined typical alignment columns. Each template alignment column is represented by the most similar entry in this set, reducing the template profiles to sequences over a different alphabet. The query alignment is converted to a profile of similarities of its columns to each of the typical columns, i. e. an alignment with N columns becomes a profile with $N \times 219$ entries. Thus, the filtering step is fast due to the change from comparing profiles to each other to comparing a query profile to a template sequence, both over the same 219 character alphabet. The pairwise similarities of the typical columns are pre-computed and stored so they become instant look-ups, improving runtime further.

5.2.2 *Detection of recently amplified TPRs*

We applied a multi-step procedure for the detection of TPRs recently amplified from single, non-repetitive motif instances (figure 5.3). We started from Pfam family TPR_1 (PF00515), which represents a single TPR $\alpha\alpha$-hairpin. This rather focused and restricted view of the TPR sequence space was widened with an HHblits search (section 5.2.1) with the TPR_1 *seed alignment* as query against an nr20 sequence database on the MPI Bioinformatics Toolkit (performed on 23rd of May 2013; Biegert *et al.*, 2006). Matches with E-values worse than $1e^{-3}$ were filtered out to crudely remove false positives yet maintain a good coverage of TPR-like sequences. The alignment of the remaining matches can be considered a reasonable representation of the TPR sequence space.

To detect recently amplified repeats in the complete nr database, we converted the HHblits match alignment to a HMMER profile HMM and used it as query against nr with *hmmsearch* of the HMMER3 package (Eddy, 2011). This resulted in 161316 matches in 76817 proteins. Next, all matches with less than 80% coverage of the profile HMM were omitted, i. e. we require matches to cover at least 28 of 34 TPR residues; this reduced the number of matches to 113421 in 59637 proteins. Finally, we added 102 flanking residues on both sides of each match, a number derived as three times the length of a TPR, resulting in sequence segments of approximately seven times the length of a TPR. In sequences with less than 102 residues before or after the match we simply used all available ones.

To find segments comprising highly similar repeats, we used the TRUST program (no version number available) to detect repeats and discarded non-repetitive segments (Szklarczyk and Heringa, 2004). As a side effect of this step, the aforementioned loose filtering becomes more restrictive as many false positive matches will not comprise TPR-like repeats. For the repetitive segments, TRUST returns a repeat alignment which we scored with al2co (no version number available), an algorithm that computes a conservation score between 0 and 9 for every alignment column (Pei and Grishin, 2001). We stringently filtered the alignments to those with averaged column conservation score above 7.

Assuming that these segments represent the currently available complement of highly similar, repetitive TPR sequences, the next step was to assess whether these repeats were homologous to singleton TPR instances. To this end, the central TPR-like sequence detected by HMMER represented its segment as query in a BLAST (version 2.2.22) search against nr (as available on the MPI Bioinformatics Toolkit on May 23rd 2013). Again, TRUST and al2co were used to detect and score repeats in the resulting matches. Unlike before, non-repetitive matches and repetitive matches with conservation scores below 5—a

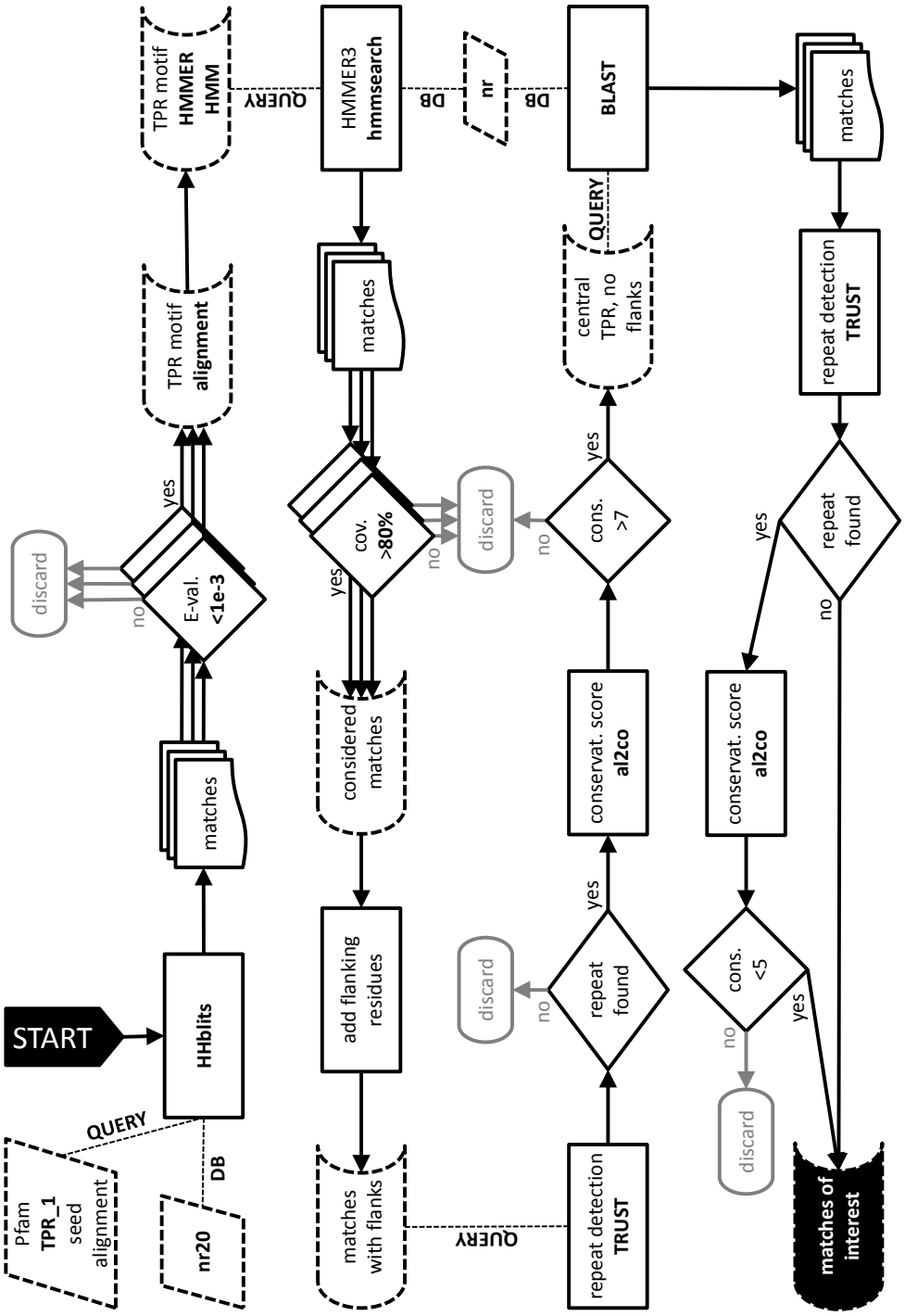


Figure 5.3: Flow diagram of the approach to detect recently amplified repeats. See figure 4.2 for explanations of the used shapes.

potential sign of false positive repeat prediction—were the target and we detected 361 candidate matches with these properties.

5.2.3 *Dataset of TPR-like αα-hairpins*

We derived a dataset depicting sequence and structure space around single, TPR-like αα-hairpins (figure 5.4). The algorithmic parameters used during the creation of this dataset were chosen empirically. The general rationale was to rather include false positives and filter them out later than to miss matches. First, all αα-hairpins with high structural similarity to single TPR instances were collected. Next, these matches became queries in homology searches. Finally, the pairwise similarities of the entries in the dataset were used for clustering.

We applied a narrow initial definition of TPR αα-hairpins by only considering the structures in the Pfam (release 27.0) family TPR_1 (PF00515), which represents a single TPR αα-hairpin. After removing structures with less than 30 structurally resolved residues in the TPR

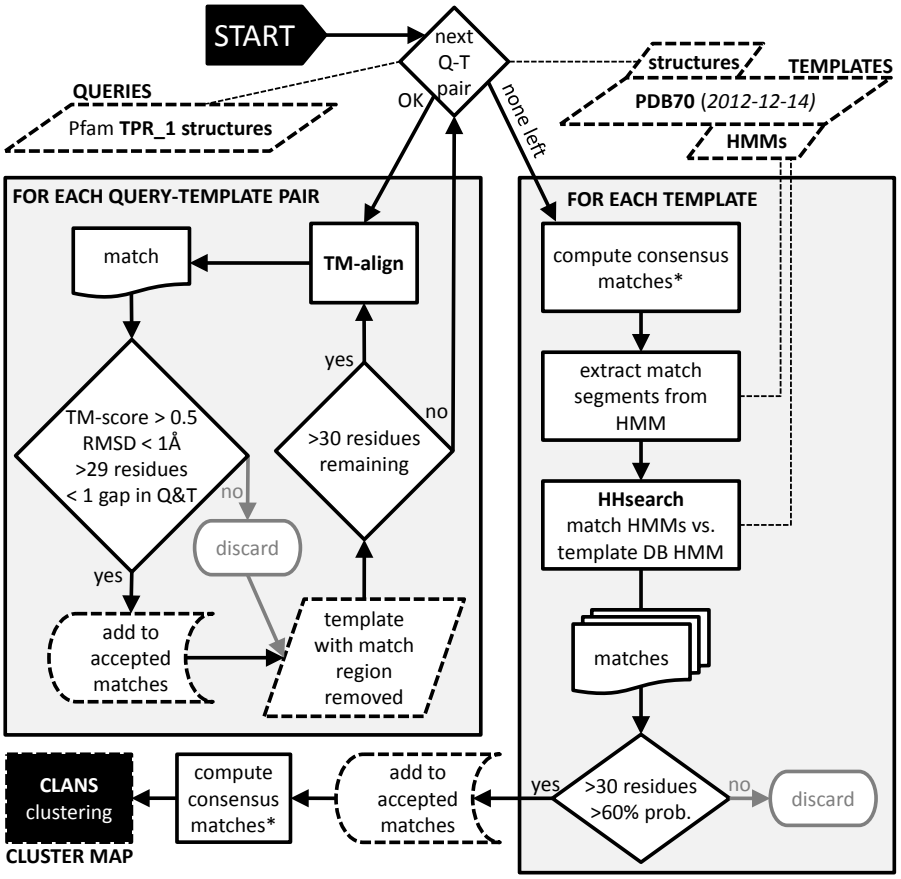


Figure 5.4: Flow diagram of our TPR-like αα-hairpin dataset creation approach. The consensus computation steps marked with asterisks are detailed in the main text. See figure 4.2 for explanations of the used shapes.

region, 106 structures were left. We used them as queries in a structural alignment procedure which we designed to detect TPR-like $\alpha\alpha$-hairpins in templates from a PDB70 database (as available on the 14th of December 2013). This database was obtained from the HHpred web server and included profile HMMs for all entries, which we planned to use in the homology-based extension. To find as many TPR-like segments of a structure as possible, each query was repeatedly aligned to each template using TM-align (version 2012/05/07; Zhang and Skolnick, 2005).[1] After each iteration, the matching region was deleted from the template structure, making it smaller every time. Our initial assumption of the first encounter with a match below a threshold TM-score as stopping criterion turned out to miss high-quality matches. The reason was the heuristic initialization used in TM-align, which is prone to find local optima (see section 2.4). We instead searched all templates completely by stopping as soon as less than 30 residues remained. Matches with TM-score above 0.5, RMSD below 1Å, more than 29 residues, and maximally one gap in both query and template were retained.

Many different queries had similar or identical matches to the same template, thus we created consensus matches by binning overlapping matches. The longest match against a template was assigned a new bin. The succeeding matches in order of decreasing length were added to an existing bin, if the overlap with any sequence already in the bin was higher than 80% of the average sequence length in that bin. Otherwise, the match spawned a new bin. To avoid indefinite extension, we limited the total sequence length covered by one bin to 50 residues. Finally, the sequence covered by each bin was computed and used from here on. This concluded the initial structure-based dataset extension step.

The second dataset extension step was conducted using HHsearch homology searches. For each result sequence from the structure-based steps, we used CSB (Kalev *et al.*, 2012) to extract the corresponding profile HMM segments from the full-length profile HMM in the aforementioned PDB70. The profile HMMs were used as query in HHsearch searches against the same PDB70 database. We chose custom parameters for a maximum of 20 matches per query-template pair and global alignment to allow for more matches in each template and higher query profile HMM coverage, respectively. Matches with more than 30 residues and an HHsearch probability above 60% were retained and overlaps stemming from matches of different queries to the same template were resolved as described above.

For the resulting sequences, we again extracted the corresponding profile HMM segment. To assess the pairwise similarity between these

1 TM-align is limited to the first chain of its input PDB files. This chain must comprise at least 4 usable residues, where usable means a structurally resolved C_α atom without alternative location indicator or with indicator *A*. We ignored structures violating any of these prerequisites.

profile HMMs, we used each one as HHsearch query against a database comprising them all. HHsearch parameters were chosen as before with the addition of ignoring secondary structure in the scoring. To analyze the dataset, a CLANS (section 2.3) cluster map was created using the pairwise p-values from these searches as similarities. This means that we represented each sequence in the cluster map with its profile HMM for the sake of more sensitive scoring.

5.3 RESULTS

5.3.1 *Recently amplified TPRs*

It is known for other proteins, e. g., β-propellers, that the amplification of whole domains from single motif instances is an ongoing process (Chaudhuri *et al.*, 2008). As we discussed before, sequence similarity is the hallmark of homology as is quickly decreases due to divergence. A reasonable approach to detect recently amplified repeats thus is to search for domains in which all repeats have completely or almost identical sequences. We performed such an analysis to evaluate whether TPR domains with a recent origin from singleton TPRs could be found.

In a first step, we detected many proteins with highly similar repeats (section 5.2.2). For each of these proteins, we determined if a single motif instance with high sequence similarity to the detected repeats could be found in another, non-repetitive protein. We detected 361 potential singletons and re-analyzed them with the sophisticated repeat detection program HHrepID (Biegert and Söding, 2008). The HHrepID results revealed all potential singletons as repetitive. The discrepancy between the initial proposal as single instances versus the final assessment as repetitions can be attributed to the higher sensitivity of HHrepID over the initially used TRUST algorithm, which we chose for its speed (section 5.2.2; Szklarczyk and Heringa, 2004).

5.3.2 *Evolution of TPR $\alpha\alpha$-hairpins*

To gain an overview over the TPR sequence space, we derived a comprehensive dataset of TPR-like sequences (section 5.2.3). Following the aforementioned assumption of a singleton ancestral form of TPRs, our dataset was centered around single $\alpha\alpha$-hairpins. The drawback of this approach is a high number of expected false positives and negatives in homology searches due to the short length of the TPR motif. To compensate for false negatives, we enriched the dataset using structure-based searches, which had the negative side-effect of introducing an analogous criterion but was necessary for better coverage of TPR sequence space. This step resulted in many singletons and disconnected *islands* of analogous matches with no obvious connection to the other

motifs. Next, the dataset was expanded using homology searches. Besides its main purpose of adding TPR homologs, this step was considered useful for the detection of intermediates between TPRs and the aforementioned sequence islands, which might reveal an otherwise occluded homology. Finally, we derived a cluster map from the pairwise similarities of all database entries to reveal false positives and ultimately analyze the dataset.

To determine a suitable cutoff value for the inclusion of similarity p-values, we analyzed the complete cluster map starting with the stringent cutoff $1e^{-10}$ (figure 5.5). At this threshold, we annotated clusters connected to the main TPR-like group and then loosened it by one order of magnitude. We iterated this procedure and found that a cluster of clearly unrelated matches connected strongly to TPR-like proteins. At a cutoff of $1e^{-3}$, the DHDPS family of TIM barrel fold proteins became connected to the TPRs. The connections are based on matches to $\alpha\alpha$-hairpins formed by the terminal TIM barrel α-helix and the initial helix of a succeeding helical domain. Due to this clear false-positive chance match, we choose $1e^{-4}$ as cutoff. The following discussions are based on a focused cluster map created by extracting all sequences connected to the main TPR hub at this cutoff (figure 5.6).

Different TPR variants and homologs cluster in one hub of the map and are mostly hard to resolve, even at more stringent cutoffs (figure 5.6). Almost all proteins are from Pfam clan TPR CL0020 and the remaining sequences either cannot be classified or are part of clan-less families. Notable examples of proteins in the hub are the membrane vesicle-coating clathrin (PF00637) and the mitochondrial protein import presequence receptor TOM20 (PF06552; Ybe *et al.*, 1999; Perry *et al.*, 2006) Clathrin repeats are shorter than TPRs, lack their distinct hydrophobic residue spacing, and adopt a unique straight superhelical fold (Ybe *et al.*, 1999). In contrast, translocase of outer membrane 20kDa subunit (TOM20) has insertions in both repeat helices, yet resembles TPR domains with its binding site on the concave side of the superhelix (Perry *et al.*, 2006). With these properties, clathrin and TOM20 can be considered diverse TPR homologs.

Another group that radiates from the hub are SEL1-like proteins (PF08238). This 36–44 residue motif, which is a negative regulator of LIN12/Notch developmental signal receptor proteins, is closely related to TPRs and remains connected to them at stringent cutoffs (Grant and Greenwald, 1996).

Similarly, half-a-TPR motifs (HAT; PF02184) are on the periphery of the hub, likely due to the different in its conservation pattern relative to TPRs (Preker and Keller, 1998). HAT motifs have been found in several proteins and are prominently featured as 12 repeats in the two HAT domains of cleavage stimulating factor 77 (CstF-77; Liu *et al.*, 2006; Bai *et al.*, 2007). Like TPRs, HAT domains are protein-protein in-

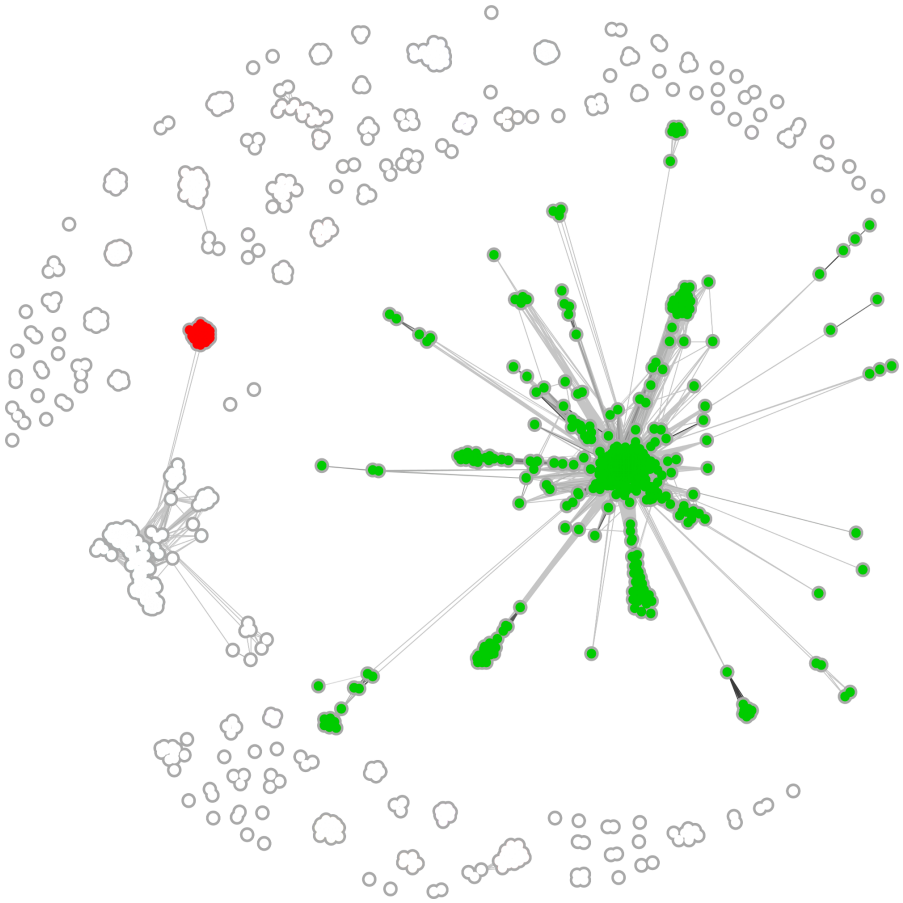


Figure 5.5: Cluster map of the TPR-like dataset. Dots represent sequences and lines depict sequence similarities based on HHsearch p-values. Clustering was iterated until equilibrium at cutoff $1e^{-4}$ and only connections below the cutoff are shown. The TPRs form the hub of a connected component (green), whereas the disconnected sequences are white. The DHDPS family (red) has false-positive connections to the TPR component at less stringent cutoffs and thus determined the cutoff value.

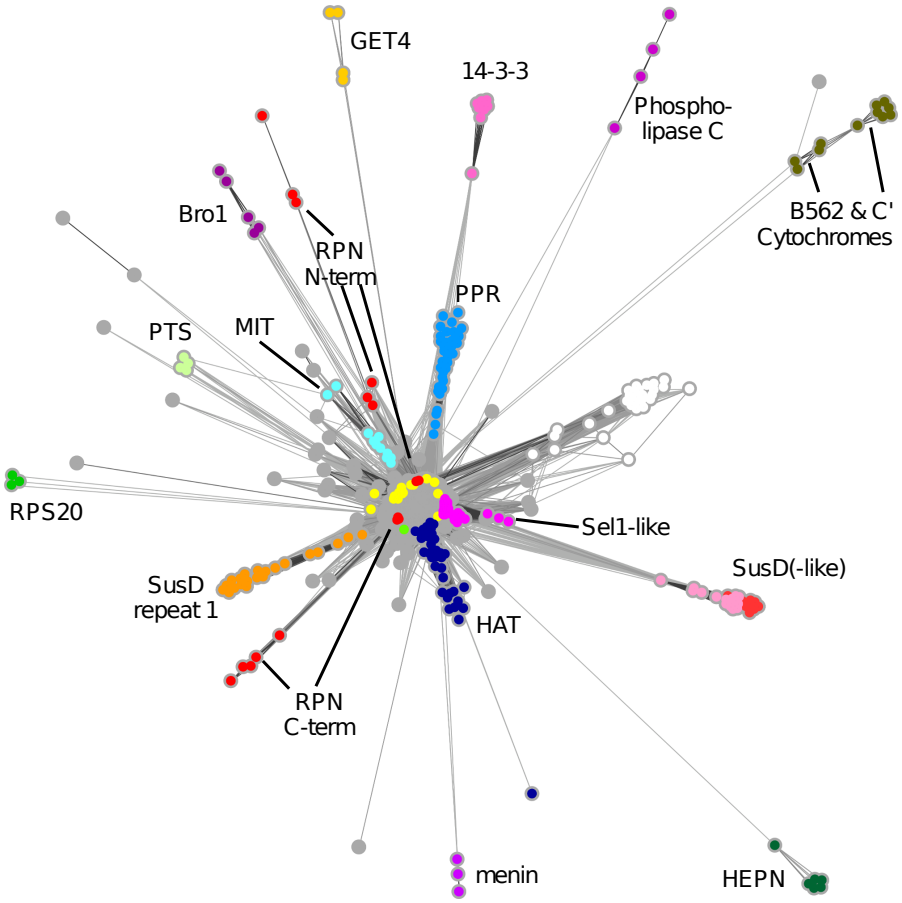


Figure 5.6: Cluster map of the connected component at cutoff $1e^{-4}$ that includes TPRs (green in figure 5.5). We reclustered the map after removing the disconnected sequences to rid the layout of the repulsive influence of these sequences. In the central hub, unlabeled sequences are: yellow = clathrin heavy chain; light green - TOM20. The unlabeled cluster of white dots comprises $\alpha\alpha$-hairpins spanning helix *B* of one TPR and helix *A* of the succeeding one; these are false positive matches of our dataset assembly process.

teraction modules, e. g., the CstF-77 HAT domains help in assembling the heterotrimeric CstF complex (Legrand *et al.*, 2007).

The outskirts of the hub are populated with clusters comprising different helical motifs. One of these clusters is formed by the microtubule interacting and trafficking domains (MIT; PF04212), a conserved sequence motif that binds the MIT interacting motif prevalent in ESCRT-III complex proteins (Scott *et al.*, 2005a; Bowers *et al.*, 2004; Katoh *et al.*, 2005; Ciccarelli *et al.*, 2003; Phillips *et al.*, 2001; Hurley and Hanson, 2010) MIT domains adopt an asymmetric three-helical bundle that resembles helices *A*-*B*-*A*′ of 1.5 TPR motifs in various structural characteristics (Scott *et al.*, 2005b). In addition, the loop between helices *A* and *B* of the MIT domain in Vps4 binds phosphoinositide in a Ca^{2+}-dependent manner (Iwaya *et al.*, 2013).

In two clusters we found matches to the 26S proteasome 19S regulatory particle non-ATPases (Rpn; PF10602) 3, 5, 6, and 7, which are part of the lid complex. The matches are hairpins with high structural similarity to TPRs, however their helices comprise an additional turn and are 40 residues long (Unverdorben *et al.*, 2014). No conserved motif was found in Rpn6 but like TPRs, these proteins are interaction modules, in this case with several 26S proteasome core particle components (Pathare *et al.*, 2012; Unverdorben *et al.*, 2014).

The cluster map also contains matches to the pentatrico peptide repeat motif (PPR; PF01535), which is considered a close TPR homolog (Small, 2000). The pentatrico peptide repeat (PPR) motif is involved in RNA editing and some PPR domains feature TPR-like N-terminal caps and C-terminal solvating helices whose function is yet unknown (Yagi *et al.*, 2013; Yin *et al.*, 2013). PPRs are limited to eukaryotes and low numbers with the exception of terrestrial plants, in which unknown causes resulted in substantial amplifications. The resulting imbalance is observable in the 70 fold higher number of PPRs in *Arabidopsis thaliana* compared to humans (Lurin *et al.*, 2004). The high diversity expected from such an amplification is reflected in the differences in strength and amount of connections of PPRs to the cluster map hub. PPR proteins can be divided into two classes based on the occurrence of canonical (*P*, 35 residues), long (*L*, 35–36 residues, often without the conserved canonical terminal proline), and short (*S*, 31 residues) motif forms (Lurin *et al.*, 2004). *P*-class PPRs comprise only canonical motifs whereas combinations of the forms are found in the plant-exclusive *PLS*-class proteins, usually in *PLS* order (Lurin *et al.*, 2004). In our cluster map, *P*- and *L*-form PPRs cluster together, as expected from their similar sequences and lengths (Lurin *et al.*, 2004). *S* forms are absent from our dataset, even though it comprises *P* and *L* forms of *PLS*-class proteins like, e. g., thylakoid assembly 8-like protein (THA8L; 4LEU), a *LPPPS* PPR (Ban *et al.*, 2013). We confirmed with TPRpred that the THA8L *S* form, but also the *L* form, are not easily identified as PPRs from sequence alone.

More distant from the hub are Bro1 domains (PF03097), which comprise a TPR domain-like substructure with three $\alpha\alpha$-hairpins (Kim *et al.*, 2005). Compared to TPRs, helices in Bro1 $\alpha\alpha$-hairpins are longer (22–30 residues). In both TPR domains and Bro1, ligands are bound similarly in the concave pocket and in Bro1 ligand binding improves stability significantly (Kim *et al.*, 2005; Scheufler *et al.*, 2000).

Two further branches comprise different $\alpha\alpha$-hairpins from starch utilization system protein D (SusD; Pfam families PF07980, PF12741, PF12771, and PF14322), a protein in *Bacteroidetes*glycan metabolism (Koropatkin *et al.*, 2008). SusD is part of the Sus system in which it contributes to oligosaccharide binding with a surface pocket and higher affinity for starch in synergy with other Sus proteins (Martens *et al.*, 2009). Of the four presumed TPR-like hairpins, the third one radiates out of the hub in the cluster map whereas matches to the fourth hairpin form a more distant outgroup. The missing first and second motif instances comprise large insertions of additional secondary structural elements in the intra-motif loop and thus lack the distinct TPRs hairpin structure questioning their labeling as TPRs. While function of the TPR-like region of SusD is unknown an involvement in SusCDEFG complex formation was discussed (Koropatkin *et al.*, 2008).

A loosely connected outgroup is formed by 14-3-3 proteins, which are expressed by all eukaryotic cells (PF00244; Obsil and Obsilova, 2011). Their nine α-helices adopt a $\alpha\alpha$-hairpin superhelix similar to TPR domains. The functional form of 14-3-3 domains is a head-to-head dimer in which the conserved residues on the concave superhelix sides form a *U*-shaped binding site (Obsil and Obsilova, 2011). Once ligand(s) are bound, 14-3-3 proteins are able to directly modify them, restrict competing access to them, or scaffold a complex with additional ligands (Fu *et al.*, 2000; Obsil *et al.*, 2001; Tzivion and Avruch, 2002; Bridges and Moorhead, 2005; Aitken *et al.*, 2002).

The aforementioned RPS20, which has been discussed as a potential ancestor of the TPR hairpin (Karpenahalli, 2006), is also in our cluster map as a loosely connected outgroup. Its connections to the hub are lost at slightly more stringent cutoffs, making the RPS20 proteins an isolated cluster. It is currently unclear why the signal between RPS20 and the main hub is comparatively weak.

Motifs from cytochromes *B*562 and *C′* are connected to the hub at a similar distance as RPS20. Indeed, there are few and weak connections between *B*562 and RPS20 but not to cytochrome *C′*, which also is not connected to the hub itself but transitively via *B*562. Both cytochromes are four-helical bundles with a central heme group that is covalently bound in cytochrome *C′* but not *B*562 (Finzel *et al.*, 1985; Arnesano *et al.*, 1999). We computed sequence alignments of the cytochrome *B*562 matches and the TPR consensus sequence that revealed a lack of consensus residues in the former except for *Gly*15

(figure 5.2), which is common in this position of turns and not specific to TPRs.

Another small cluster is formed by menin, which is a putative suppressor of multiple endocrine neoplasia type 1 (MEN1; Thakker, 2010). Besides its anti-tumor function, menin binds many more proteins and affects a diverse set of cellular processes (Busygina *et al.*, 2006; Yang and Hua, 2007; Chen *et al.*, 2008). Menin adopts four distinct domains, including a superhelical domain with three $\alpha\alpha$-hairpins that is TPR-like in structure and sequence (Huang *et al.*, 2012). Like in TPR domains, the concave surface of the superhelix contributes to ligand binding however a neighboring domain is also necessary for proper binding.

Further, one cluster contains enzyme IIA domain variants of the prokaryotic phosphotransferase system (PTS; Tang *et al.*, 2005). The PTS is a signal transduction network that translocates sugars across the cytoplasmic membrane (Robillard and Broos, 1999). Enzyme II is the sugar-specific component of PTS and it comprises three domains of which the first, IIA, is a homotrimeric three-helical bundle. The matches in our cluster map are to lactose- and cellobiose-specific IIA domains. The region covered by each match are the two N-terminal helices and the $\alpha\alpha$-hairpin they adopt can be superimposed well onto canonical TPRs motif structures (RMSD ~1Å). Not many residues of these two helices adhere to the TPR consensus but they share a similar pattern of hydrophobic residues. The C-terminal, third helix is unsurprisingly dissimilar given its structural role during trimerization where it forms a coiled-coil (Sliz *et al.*, 1997).

Another cluster comprises $\alpha\alpha$-hairpins from guided-entry of tail-anchored proteins (Get) pathway protein Get4, which is involved in targeting tail-anchored proteins to the endoplasmic reticulum membrane for insertion (Bozkurt *et al.*, 2010). Get4 adopts a α-solenoid structure with 7 $\alpha\alpha$-hairpins and its structural similarity to TPR has been noted including presumed ligand binding on the convex surface. Due to unconserved lengths for hairpin α-helices and the less pronounced superhelical twist, a common ancestry with TPRs was found to be debatable (Chang *et al.*, 2010). The connections in our cluster map, accompanied by TPRpred analyzes, revealed the hairpin formed by the second and third α-helices as TPR-like, however with weak sequence similarity. The length diversity observed in the other $\alpha\alpha$-hairpins hinders their classification and we could not detect repeats from sequence using HHrepID. With these findings, it remains unclear whether the repeats are highly divergent TPRs or not homologous to TPRs but similar in one hairpin by chance.

Similar to the aforementioned Get4 proteins, a single TPR $\alpha\alpha$-hairpin connects phospholipase C (α-toxin) from *Clostridium absonum* to the main hub. Unlike TPRs, the N-terminal domain adopts a globular fold and the TPR-like $\alpha\alpha$-hairpin is located on the surface towards a β-sheet

of the C-terminal β-sandwhich domain, with which it might interact (Clark *et al.*, 2003). We were able to confirm the TPR-likeness of the matching region with TPRpred, however with low probability (E-value $2.7e^{-2}$). Overall, it seems unlikely that the $\alpha\alpha$-hairpin evolved from a TPR motif as only a single protein of this family connects to the central hub of our cluster map and has very low similarity scores—the other cluster members are daisy-chained at the chosen cutoff and the cluster dissolves and becomes disconnected from the hub at stricter cutoffs.

The last cluster comprises higher eukaryotes and prokaryotes nucleotide-binding domains (HEPN; PF05168), which are found in multidomain proteins of vertebrates as well as in single and two-domain proteins of bacteria and archaea (Grynberg *et al.*, 2003). Mutations in the HEPN domain of human sacsin cause the early-onset neurodegenerative disease autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS; Kozlov *et al.*, 2011) HEPN domains adopt a four-helical up-and-down bundle fold and bind guanosine triphosphate (GTP) with high affinity (Kozlov *et al.*, 2011; Erlandsen *et al.*, 2004). The first $\alpha\alpha$-hairpin is structurally similar to a TPR motif, however we could only detect a conserved pattern of hydrophobics and no TPR consensus residues on the sequence level, reminiscent of the aforementioned PTS enzyme IIA domains.

5.4 CONCLUSIONS

With the search for repeats that were recently amplified from singletons and the overview provided by our cluster map, we analyzed two different aspects of TPR evolution. As we were unable to detect TPR domains recently amplified from singletons without repetitive context, it seems likely that this is not an ongoing process. Instead, modern TPR domains are probably the result of duplications of fully formed domains with later adaptations to their new environment.

In our analysis of the TPR-centric cluster map, we were able to rediscover known relationship like the ones to PPR and SEL1-like motifs. The *PLS* class of PPRs proteins is noteworthy as the *L* and *P* motif forms of a protein might be part of our cluster map even if the similar yet slightly shorter *S* form found in the same protein is not. Our dataset assembly procedure has no obvious restrictions that would cause such exclusions and a detailed analysis of the differences between the forms would be interesting. Even though these details are beyond the scope of this work, a key point might be the sensitivity of homology detection procedures for short sequences.

Homology searches with short fragments, e. g., the 34 residues of TPR, are likely to cause a large number of false positives and negatives. We found in our complete cluster map that many matches connected to TPR at reasonable cutoffs but were far off the main hub (figure 5.5).

Indeed, we consider many of these connections as false positives that are not indicative of a homologous origin. However, this also questions the use of contemporary homology detection methods for very short sequences. Findings based on such weak links should be carefully evaluated with additional methods and ideally manual expert analyzes.

Interestingly, false positives did not seem to be a pronounced problem in the motifs closely surrounding the TPR hub in our cluster map. Almost all of these motifs radiate from the main hub as distinct groups with few or no connections to neighboring clusters. This similarity pattern presumably reflects single specialization events that branched the motif off one TPR hairpin or domain. Later adaptations would then only affect one group of homologs and each amino acid alteration would lower the similarity between previously close motifs.

While this course of evolution is often observed especially when evolving from well established stable motifs, other possibilities exist. In our cluster map we found cytochromes *B*562 and *C′* as a case of daisy-chained motif derivation. While cytochrome *B*562 is connected to the hub directly and likely derived from TPRs, cytochrome *C′* has merely transitive connections to the hub via *B*562 and should be considered a derivative thereof. Notably, this is the only such case we found.

Overall, the 34 residues of TPRs seem to be near the border of reliable automatic detection of remote homologs. Alternative approaches to pairwise profile HMM comparisons exist and especially Markov random fields (MRFs) have increased in popularity recently in various areas of protein bioinformatics (Gehrmann *et al.*, 2013; Daniels *et al.*, 2012; Menke *et al.*, 2010; Li *et al.*, 2007). As apparent from the available results, proteins with high β-strand content benefit most from MRFs, which are capable of capturing the interdependencies of non-neighboring residues. But the application of MRFs to bioinformatic problems is often difficult due to the high amount of required data and computational resources and their full potential has yet to be determined (Daniels *et al.*, 2012).

6 OUTLOOK

In this work, we presented three case studies of remote protein homologies based on whole domains as well as on large and small supersecondary structural elements. The application of state-of-the-art homology detection and sequence clustering methods provided us with high-quality data from which we derived structure and function predictions as well as evolutionary scenarios.

First, we analyzed the evolution of the SMP domain, before merely known as sequence motif in several proteins but without structural or functional description. It became evident that SMP domains are part of a superfamily that also includes BPI-like and Takeout-like domains and which we named the tubular lipid-binding proteins (TULIP) domain superfamily. Based on this assessment, we proposed that SMP domains adopt a TULIP domain fold and bind hydrophobic ligands. A recent study confirmed these predictions with the presentation of the first SMP domain structure, which indeed adopts a TULIP domain fold and binds lipids in its central tunnel. However, it remains elusive how SMP domains transport lipids between the membranes, providing an important and interesting avenue for future experimental studies.

Next, we investigated whether β-propeller blades also gave rise to other folds. Using only sequence-based remote homology detection, we detected four different folds with potential homology to β-propeller blades. To differentiate between homology and structure-induced similarity, we analyzed the relationship of changes in structure and sequence for presumably homologous regions. Our results indicate that the blade-matching β-meanders of type II β-prisms (BP2s) and inositol-requiring enzyme 1 luminal domains (IRE1-LDs) are actual blade homologs, whereas those of WW domains and β-pinwheels are an analogous development. To substantiate our claim of an homologous origin of β-propellers and BP2 proteins, we are currently examining if a designed protein with three repeats of a PQQ β-propeller blade can adopt a BP2 fold with no or only minor modifications.

Finally, we provided an overview of the sequence space surrounding the tetratrico peptide repeat (TPR) motif. Our search for recently amplified TPR domains revealed that novel TPR domains are today only formed by duplication and subsequent modification of complete domains. While this contrasts the ongoing evolution of other repeat proteins like β-propellers, the adaptation of pre-established stable scaffold to new functions is not surprising either. The cluster map of TPR-like motifs shows that many $\alpha\alpha$-hairpin motifs are closely related to TPRs. Similar to TPR domains, these motifs are often found in super-

helical arrangements relevant for protein-protein interaction or ligand binding. To elucidate the exact relationship within this tight cluster of TPR-like proteins would require thorough case-by-case analyzes, which was beyond the scope of this work. However, the evolutionary trajectory resulting from such an analysis including the changes that led to the very different binding specificities of these motifs would be highly interesting.

While these three projects highlight the possibilities of current approaches to remote homology detection, method development has not halted in this area. As we mentioned before, Markov random fields (MRFs) generalize hidden Markov models (HMMs) to dependencies between arbitrary states and are thus a logical improvement over the pairwise profile HMM comparisons used for homology assessment throughout this work. However, because MRFs require more training data for model creation and have longer model comparison runtime, they are currently not readily applicable in many cases. The expected improvement of protein database depth, MRF implementations, and computational resources in the coming years might remedy these drawbacks completely, making MRF the new standard for remote homology detection. Overall, the required expertise and the necessary amount of time for studies like the ones presented here shows that further methodological advancements are necessary before non-experts can benefit from very remote homologies.

Similarly, the clustering approach in CLANS might need to be evaluated against newer methods. The large number of successful studies using it show that it is clearly useful for protein sequence analyzes, but the underlying Fruchterman-Reingold algorithm also has its problems, e. g., a high risk of finding local minima and long runtime. Many dimensionality reduction algorithms exist and it would be interesting to evaluate their suitability for protein clustering (see, e. g., van der Maaten and Hinton, 2008; Shieh *et al.*, 2011). A thorough comparison and the inclusion of the most suitable algorithm(s) in CLANS would surely be beneficial and should be pursued in the near future.

To summarize, this work contributes to the knowledge base on remote homology on the structural level of domains and supersecondary structural elements (super-SSEs). Probing the limits of contemporary approaches, our findings also highlight their inherent possibilities for knowledge transfer and evolutionary studies. We were able to show that the very remote homology between SMP and bactericidal/permeability-increasing proteins (BPI)-like domains is sufficient to provide reliable function and structure predictions. The reconstructed history of TULIP domains shows that it might be feasible to establish detailed evolutionary scenarios for even very remotely related domains and possibly all extant proteins, given the limited number of folds in nature. The available remote homology detec-

tion methods will uncover many more interesting cases of remote and fold-spanning homologies, similar to our detection of other folds based on β-propeller blades. Further, novel algorithms might provide even more insights into currently difficult cases like the TPR repeats. And, ultimately, these results connect entities otherwise separate in protein classifications and could one day break the ground for a classification by natural descent.

A

PUBLICATIONS

Some ideas and figures throughout this work have previously appeared in the following publications:

- *Kopec K.O.* and Lupas A.N. *β-propeller Blades as Ancestral Peptides in Protein Evolution.* PLoS ONE, *8*(10): e77074, 2013.
  - doi:10.1371/journal.pone.0077074
  - Content from this publication appears mostly in chapter 4.
  - Conceived and designed the experiments: *KOK* ANL. Performed the experiments: *KOK*. Analyzed the data: *KOK* ANL. Wrote the paper: *KOK* ANL.
- Kalev I., Mechelke M., *Kopec K.O.*, Holder T., Carstens S., and Habeck M. *CSB: a Python framework for structural bioinformatics.* Bioinformatics, *28*(22): 2996-2997, 2012.
  - doi:10.1093/bioinformatics/bts538
  - `https://csb.codeplex.com`
  - Designed the software, tested its performance, and wrote the paper: IK. Contributed modules and test cases: IK MM *KOK* TH SC MH.
- *Kopec K.O.*, Alva V., and Lupas A.N. *Bioinformatics of the TULIP domain superfamily.* Biochemical Society transactions, *39*(4): 1033-1038, 2011.
  - doi:10.1042/BST0391033
  - Content from this publication appears mostly in chapter 3.
  - Conceived and designed the experiments: *KOK* VA ANL. Performed the experiments: *KOK* VA. Analyzed the data: *KOK* VA ANL. Wrote the paper: *KOK* VA ANL.
- *Kopec K.O.*, Alva V., and Lupas A.N. *Homology of SMP domains to the TULIP superfamily of lipid-binding proteins provides a structural basis for lipid exchange between ER and mitochondria.* Bioinformatics, *26*(16): 1927-1931, 2010.
  - doi:10.1093/bioinformatics/btq326
  - Content from this publication appears mostly in chapter 3.
  - Conceived and designed the experiments: *KOK* VA ANL. Performed the experiments: *KOK* VA. Analyzed the data: *KOK* VA ANL. Wrote the paper: *KOK* VA ANL.

B

CONTRIBUTIONS

A significant amount of ideas for the research in this work was developed in discussions with Prof. Dr. Andrei N. Lupas and Dr. Vikram Alva (VA). Most single ideas are not traceable to one person any more as they developed over time, but it is noteworthy that this dissertation is not conceivable without their contributions. In addition, VA and Jens Baßler read drafts of the dissertation and suggested improvements.

CHAPTER 2

Figures 2.1, 2.2, and 2.3 are in the public domain and thus no permission is necessary.

CHAPTER 3

This work is part of two manuscripts that have been previously published in Bioinformatics (Kopec *et al.*, 2010) and the Biochemical Society transactions (Kopec *et al.*, 2011). Text and figures from these manuscripts have been reproduced with permission from the publishers. Both projects were carried out in collaboration with Vikram Alva (VA) and Andrei N. Lupas (ANL). ANL initially recognized the similarity between SMP domains and BPI. I, VA, and ANL conceived and designed the following experiments. I and VA performed the experiments. I, VA, and ANL analyzed the data and wrote the paper. I and VA created the figures.

CHAPTER 4

This work is part of a manuscript that has been previously published in PLoS ONE (Kopec and Lupas, 2013). Text and figures from this manuscript have been reproduced with permission from the publisher. The project was carried out in collaboration with Andrei N. Lupas (ANL). I and ANL conceived and designed the experiments. I performed the experiments. I and ANL analyzed the data and wrote the paper. I created the figures.

CHAPTER 5

This work was carried out in collaboration with Andrei N. Lupas (ANL). I and ANL conceived and designed the experiments. I performed the experiments. I and ANL analyzed the data. I created the figures.

Achleitner G., Gaigg B., Krasser A., Kainersdorfer E., Kohlwein S.D., Perktold A., Zellnig G., and Daum G. *Association between the endoplasmic reticulum and mitochondria of yeast facilitates interorganelle transport of phospholipids through membrane contact*. European Journal of Biochemistry, *264*(2):545–553, 1999. doi:10.1046/j.1432-1327.1999.00658.x.

Aitken A., Baxter H., Dubois T., Clokie S., Mackie S., Mitchell K., Peden A., and Zemlickova E. *Specificity of 14-3-3 isoform dimer interactions and phosphorylation.* Biochemical Society transactions, *30*(4):351–60, 2002.

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. *Basic local alignment search tool.* Journal of molecular biology, *215*(3):403–10, 1990. doi:10.1016/S0022-2836(05)80360-2.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., and Lipman D.J. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic acids research, *25*(17):3389–402, 1997. doi:10.1093/nar/25.17.3389.

Alva V., Ammelburg M., Söding J., and Lupas A.N. *On the origin of the histone fold.* BMC structural biology, *7*:17, 2007. doi:10.1186/1472-6807-7-17.

Alva V., Koretke K.K., Coles M., and Lupas A.N. *Cradle-loop barrels and the concept of metafolds in protein classification by natural descent.* Current opinion in structural biology, *18*(3):358–65, 2008. doi:10.1016/j.sbi.2008.02.006.

Alva V., Remmert M., Biegert A., Lupas A.N., and Söding J. *A galaxy of folds.* Protein science : a publication of the Protein Society, *19*(1):124–30, 2010. doi:10.1002/pro.297.

Andreeva A., Howorth D., Chandonia J.M., Brenner S.E., Hubbard T.J.P., Chothia C., and Murzin A.G. *Data growth and its impact on the SCOP database: New developments*. Nucleic Acids Research, *36*, 2008. doi:10.1093/nar/gkm993.

Andreeva A. and Murzin A.G. *Evolution of protein fold in the presence of functional constraints.* Current opinion in structural biology, *16*(3):399–408, 2006. doi:10.1016/j.sbi.2006.04.003.

Andreeva A., Prlić A., Hubbard T.J.P., and Murzin A.G. *SISYPHUS–structural alignments for proteins with non-trivial relationships.* Nucleic acids research, *35*(Database issue):D253–9, 2007. doi:10.1093/nar/gkl746.

Apic G., Gough J., and Teichmann S.A. *Domain combinations in archaeal, eubacterial and eukaryotic proteomes.* Journal of molecular biology, *310*(2):311–25, 2001. doi:10.1006/jmbi.2001.4776.

Arnesano F., Banci L., Bertini I., Faraone-Mennella J., Rosato A., Barker P.D., and Fersht A.R. *The solution structure of oxidized Escherichia coli cytochrome b562.* Biochemistry, *38*(27):8657–70, 1999. doi:10.1021/bi982785f.

Bai Y., Auperin T.C., Chou C.Y., Chang G.G., Manley J.L., and Tong L. *Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors.* Molecular cell, *25*(6):863–75, 2007. doi:10.1016/j.molcel.2007.01.034.

Ban T., Ke J., Chen R., Gu X., Tan M.H.E., Zhou X.E., Kang Y., Melcher K., Zhu J.K., and Xu H.E. *Structure of a PLS-class pentatricopeptide repeat protein provides insights into mechanism of RNA recognition.* The Journal of biological chemistry, *288*(44):31540–8, 2013. doi:10.1074/jbc.M113.496828.

Barker A.R., Wickstead B., Gluenz E., and Gull K. *Bioinformatic insights to the ESAG5 and GRESAG5 gene families in kinetoplastid parasites.* Molecular and biochemical parasitology, *162*(2):112–22, 2008. doi:10.1016/j.molbiopara.2008.08.003.

Baum L.E. *An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes*. Inequalities, *3*:1–8, 1972.

Beamer L.J., Carroll S.F., and Eisenberg D. *Crystal structure of human BPI and two bound phospholipids at 2.4 angstrom resolution.* Science (New York, N.Y.), *276*(5320):1861–4, 1997. doi:10.1126/science.276.5320.1861.

Biegert A., Mayer C., Remmert M., Söding J., and Lupas A.N. *The MPI Bioinformatics Toolkit for protein sequence analysis.* Nucleic acids research, *34*(Web Server issue):W335–9, 2006. doi:10.1093/nar/gkl217.

Biegert A. and Söding J. *De novo identification of highly diverged protein repeats by probabilistic consistency.* Bioinformatics (Oxford, England), *24*(6):807–14, 2008. doi:10.1093/bioinformatics/btn039.

Bingle C.D. and Craven C.J. *PLUNC: a novel family of candidate host defence proteins expressed in the upper airways and nasopharynx.* Human molecular genetics, *11*(8):937–43, 2002. doi:10.1093/hmg/11.8.937.

Björklund A.K., Ekman D., Light S., Frey-Skött J., and Elofsson A. *Domain rearrangements in protein evolution.* Journal of molecular biology, *353*(4):911–23, 2005. doi:10.1016/j.jmb.2005.08.067.

Blatch G.L. and Lässle M. *The tetratricopeptide repeat: a structural motif mediating protein-protein interactions.* BioEssays : news and reviews in molecular, cellular and developmental biology, *21*(11):932–9, 1999.

Bork P. and Sudol M. *The WW domain: a signalling site in dystrophin?* Trends in biochemical sciences, *19*(12):531–3, 1994. doi:10.1016/0968-0004(94)90053-1.

Bowers K., Lottridge J., Helliwell S.B., Goldthwaite L.M., Luzio J.P., and Stevens T.H. *Protein-protein interactions of ESCRT complexes in the yeast Saccharomyces cerevisiae.* Traffic (Copenhagen, Denmark), *5*(3):194–210, 2004. doi:10.1111/j.1600-0854.2004.00169.x.

Bozkurt G., Wild K., Amlacher S., Hurt E., Dobberstein B., and Sinning I. *The structure of Get4 reveals an alpha-solenoid fold adapted for multiple interactions in tail-anchored protein biogenesis.* FEBS letters, *584*(8):1509–14, 2010. doi:10.1016/j.febslet.2010.02.070.

Branden C.I. and Tooze J. *Introduction to protein structure*. Garland Publishing Inc., New York, 2 edition, 1999. ISBN 0815323050.

Bridges D. and Moorhead G.B.G. *14-3-3 proteins: a number of functions for a numbered protein.* Science's STKE : signal transduction knowledge environment, *2005*(296):re10, 2005. doi:10.1126/stke.2962005re10.

Bryngelson J.D., Onuchic J.N., Socci N.D., and Wolynes P.G. *Funnels, pathways, and the energy landscape of protein folding: a synthesis.* Proteins, *21*(3):167–95, 1995. doi:10.1002/prot.340210302.

Burley S.K., Almo S.C., Bonanno J.B., Capel M., Chance M.R., Gaasterland T., Lin D., Sali A., Studier F.W., and Swaminathan S. *Structural genomics: beyond the human genome project.* Nature genetics, *23*(2):151–7, 1999. doi:10.1038/13783.

Busygina V., Kottemann M.C., Scott K.L., Plon S.E., and Bale A.E. *Multiple endocrine neoplasia type 1 interacts with forkhead transcription factor CHES1 in DNA damage response.* Cancer research, *66*(17):8397–403, 2006. doi:10.1158/0008-5472.CAN-06-0061.

Chandonia J.M., Hon G., Walker N.S., Lo Conte L., Koehl P., Levitt M., and Brenner S.E. *The ASTRAL Compendium in 2004.* Nucleic acids research, *32*(Database issue):D189–92, 2004. doi:10.1093/nar/gkh034.

Chandra N., Ramachandraiah G., Bachhawat K., Dam T.K., Surolia A., and Vijayan M. *Crystal structure of a dimeric mannose-specific agglutinin from garlic: quaternary association and carbohydrate specificity.* Journal of molecular biology, *285*(3):1157–68, 1999. doi:10.1006/jmbi.1998.2353.

Chang Y.W., Chuang Y.C., Ho Y.C., Cheng M.Y., Sun Y.J., Hsiao C.D., and Wang C. *Crystal structure of Get4-Get5 complex and its interactions with Sgt2, Get3, and Ydj1.* The Journal of biological chemistry, *285*(13):9962–70, 2010. doi:10.1074/jbc.M109.087098.

Chaudhuri I., Söding J., and Lupas A.N. *Evolution of the beta-propeller fold.* Proteins, *71*(2):795–803, 2008. doi:10.1002/prot.21764.

Chen C.K.M., Chan N.L., and Wang A.H.J. *The many blades of the β-propeller proteins: conserved but versatile.* Trends in biochemical sciences, *36*(10):553–61, 2011. doi:10.1016/j.tibs.2011.07.004.

Chen G., A J., Wang M., Farley S., Lee L.Y., Lee L.C., and Sawicki M.P. *Menin promotes the Wnt signaling pathway in pancreatic endocrine cells.* Molecular cancer research : MCR, *6*(12):1894–907, 2008. doi:10.1158/1541-7786.MCR-07-2206.

Chothia C., Gough J., Vogel C., and Teichmann S.A. *Evolution of the protein repertoire.* Science (New York, N.Y.), *300*(5626):1701–3, 2003. doi:10.1126/science.1085371.

Chou K.C. *Prediction of tight turns and their types in proteins.* Analytical biochemistry, *286*(1):1–16, 2000. doi:10.1006/abio.2000.4757.

Ciccarelli F.D., Proukakis C., Patel H., Cross H., Azam S., Patton M.A., Bork P., and Crosby A.H. *The identification of a conserved domain in both spartin and spastin, mutated in hereditary spastic paraplegia.* Genomics, *81*(4):437–41, 2003. doi:10.1016/S0888-7543(03)00011-9.

Clark G.C., Briggs D.C., Karasawa T., Wang X., Cole A.R., Maegawa T., Jayasekera P.N., Naylor C.E., Miller J., Moss D.S., *et al. Clostridium absonum α-Toxin: New Insights into Clostridial Phospholipase C Substrate Binding and Specificity*. Journal of Molecular Biology, *333*(4):759–769, 2003. doi:10.1016/j.jmb.2003.07.016.

Coles M., Hulko M., Djuranovic S., Truffault V., Koretke K.K., Martin J., and Lupas A.N. *Common evolutionary origin of swapped-hairpin and double-psi beta barrels.* Structure (London, England : 1993), *14*(10):1489–98, 2006. doi:10.1016/j.str.2006.08.005.

Copley R.R., Russell R.B., and Ponting C.P. *Sialidase-like Asp-boxes: sequence-similar structures within different protein folds.* Protein science : a publication of the Protein Society, *10*(2):285–92, 2001. doi:10.1110/ps.31901.

Corbett K.D., Schoeffler A.J., Thomsen N.D., and Berger J.M. *The structural basis for substrate specificity in DNA topoisomerase IV.* Journal of molecular biology, *351*(3):545–61, 2005. doi:10.1016/j.jmb.2005.06.029.

Corbett K.D., Shultzaberger R.K., and Berger J.M. *The C-terminal domain of DNA gyrase A adopts a DNA-bending beta-pinwheel fold.* Proceedings of the National Academy of Sciences of the United States of America, *101*(19):7293–8, 2004. doi:10.1073/pnas.0401595101.

Coulson A.F.W. and Moult J. *A unifold, mesofold, and superfold model of protein fold use.* Proteins, *46*(1):61–71, 2002. doi:10.1002/prot.10011.

Credle J.J., Finer-Moore J.S., Papa F.R., Stroud R.M., and Walter P. *On the mechanism of sensing unfolded protein in the endoplasmic reticulum.* Proceedings of the National Academy of Sciences of the United States of America, *102*(52):18773–84, 2005. doi:10.1073/pnas.0509487102.

Csaba G., Birzele F., and Zimmer R. *Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis.* BMC structural biology, *9*:23, 2009. doi:10.1186/1472-6807-9-23.

D'Andrea L.D. and Regan L. *TPR proteins: the versatile helix.* Trends in biochemical sciences, *28*(12):655–62, 2003. doi:10.1016/j.tibs.2003.10.007.

Daniels N.M., Hosur R., Berger B., and Cowen L.J. *SMURFLite: combining simplified Markov random fields with simulated evolution improves remote homology detection for beta-structural proteins into the twilight zone.* Bioinformatics (Oxford, England), *28*(9):1216–22, 2012. doi:10.1093/bioinformatics/bts110.

Das A.K., Cohen P.W.T.W., and Barford D. *The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions.* The EMBO journal, *17*(5):1192–9, 1998. doi:10.1093/emboj/17.5.1192.

Dauwalder B., Tsujimoto S., Moss J., and Mattox W. *The Drosophila takeout gene is regulated by the somatic sex-determination pathway and affects male courtship behavior.* Genes & development, *16*(22):2879–92, 2002. doi:10.1101/gad.1010302.

De Mejía E.G. and Prisecaru V.I. *Lectins as bioactive plant proteins: a potential in cancer treatment.* Critical reviews in food science and nutrition, *45*(6):425–45, 2005. doi:10.1080/10408390591034445.

Doolittle W.F. *Phylogenetic classification and the universal tree.* Science (New York, N.Y.), *284*(5423):2124–9, 1999. doi:10.1126/science.284.5423.2124.

Durbin R., Eddy S.R., Krogh A., and Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, United Kingdom, 1 edition, 1998. ISBN 0521629713.

Eddy S.R. *What is dynamic programming?* Nature biotechnology, *22*(7):909–10, 2004. doi:10.1038/nbt0704-909.

Eddy S.R. *Accelerated Profile HMM Searches.* PLoS computational biology, *7*(10):e1002195, 2011. doi:10.1371/journal.pcbi.1002195.

Erlandsen H., Canaves J.M., Elsliger M.A., von Delft F., Brinen L.S., Dai X., Deacon A.M., Floyd R., Godzik A., Grittini C., *et al. Crystal structure of an HEPN domain protein (TM0613) from Thermotoga maritima at 1.75 A resolution.* Proteins, *54*(4):806–9, 2004. doi:10.1002/prot.10631.

Farzadfard F., Gharaei N., Pezeshk H., and Marashi S.A. *Beta-sheet capping: signals that initiate and terminate beta-sheet formation.* Journal of structural biology, *161*(1):101–10, 2008. doi:10.1016/j.jsb.2007.09.024.

Fetrow J.S. and Godzik A. *Function driven protein evolution. A possible proto-protein for the RNA-binding proteins.* Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pages 485–96, 1998.

Finn R.D., Bateman A., Clements J., Coggill P., Eberhardt R.Y., Eddy S.R., Heger A., Hetherington K., Holm L., Mistry J., *et al. Pfam: the protein families database.* Nucleic acids research, *42*(Database issue):D222–30, 2014. doi:10.1093/nar/gkt1223.

Finzel B.C., Weber P.C., Hardman K.D., and Salemme F.R. *Structure of ferricytochrome c' from Rhodospirillum molischianum at 1.67 A resolution.* Journal of molecular biology, *186*(3):627–43, 1985. doi:10.1016/0022-2836(85)90135-4.

Fitch W.M. *Distinguishing homologous from analogous proteins.* Systematic zoology, *19*(2):99–113, 1970. doi:10.2307/2412448.

Forrer P., Binz H.K., Stumpp M.T., and Plückthun A. *Consensus design of repeat proteins.* Chembiochem : a European journal of chemical biology, *5*(2):183–9, 2004. doi:10.1002/cbic.200300762.

Forrer P., Stumpp M.T., Binz H.K., and Plückthun A. *A novel strategy to design binding molecules harnessing the modular nature of repeat proteins.* FEBS letters, *539*(1-3):2–6, 2003. doi:10.1016/S0014-5793(03)00177-7.

Frickey T. and Lupas A.N. *CLANS: a Java application for visualizing protein families based on pairwise similarity.* Bioinformatics (Oxford, England), *20*(18):3702–4, 2004. doi:10.1093/bioinformatics/bth444.

Fruchterman T.M.J. and Reingold E.M. *Graph drawing by force-directed placement*. Software: Practice and Experience, *21*(11):1129–1164, 1991. doi:10.1002/spe.4380211102.

Fu H., Subramanian R.R., and Masters S.C. *14-3-3 proteins: structure, function, and regulation.* Annual review of pharmacology and toxicology, *40*:617–47, 2000. doi:10.1146/annurev.pharmtox.40.1.617.

Fujikawa K., Seno K., and Ozaki M. *A novel Takeout-like protein expressed in the taste and olfactory organs of the blowfly, Phormia regina.* The FEBS journal, *273*(18):4311–21, 2006. doi:10.1111/j.1742-4658.2006.05422.x.

Fülöp V. and Jones D.T. *β Propellers: structural rigidity and functional diversity*. Current Opinion in Structural Biology, *9*(6):715–721, 1999. doi:10.1016/S0959-440X(99)00035-4.

Gehrmann T., Loog M., Reinders M.J.T., and de Ridder D. *Pattern Recognition in Bioinformatics*, volume 7986 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-39158-3. doi:10.1007/978-3-642-39159-0.

Ghosh M., Anthony C., Harlos K., Goodwin M.G., and Blake C. *The refined structure of the quinoprotein methanol dehydrogenase from Methylobacterium extorquens at 1.94 A.* Structure (London, England : 1993), *3*(2):177–87, 1995.

Gilbert L.I., Granger N.A., and Roe R.M. *The juvenile hormones: historical facts and speculations on future research directions.* Insect biochemistry and molecular biology, *30*(8-9):617–44, 2000. doi:10.1016/S0965-1748(00)00034-5.

Grant B. and Greenwald I. *The Caenorhabditis elegans sel-1 gene, a negative regulator of lin-12 and glp-1, encodes a predicted extracellular protein.* Genetics, *143*(1):237–47, 1996.

Grishin N.V. *Fold change in evolution of protein structures.* Journal of structural biology, *134*(2-3):167–85, 2001a. doi:10.1006/jsbi.2001.4335.

Grishin N.V. *KH domain: one motif, two folds.* Nucleic acids research, *29*(3):638–43, 2001b. doi:10.1093/nar/29.3.638.

Groves M.R. and Barford D. *Topological characteristics of helical repeat proteins.* Current opinion in structural biology, *9*(3):383–9, 1999. doi:10.1016/S0959-440X(99)80052-9.

Grynberg M., Erlandsen H., and Godzik A. *HEPN: a common domain in bacterial drug resistance and human neurodegenerative proteins.* Trends in biochemical sciences, *28*(5):224–6, 2003. doi:10.1016/S0968-0004(03)00060-4.

Hamiaux C., Stanley D., Greenwood D.R., Baker E.N., and Newcomb R.D. *Crystal structure of Epiphyas postvittana takeout 1 with bound ubiquinone supports a role as ligand carriers for takeout proteins in insects.* The Journal of biological chemistry, *284*(6):3496–503, 2009. doi:10.1074/jbc.M807467200.

Henikoff S. and Henikoff J.G. *Amino acid substitution matrices from protein blocks.* Proceedings of the National Academy of Sciences of the United States of America, *89*(22):10915–9, 1992.

Hester G., Kaku H., Goldstein I.J., and Wright C.S. *Structure of mannose-specific snowdrop (Galanthus nivalis) lectin is representative of a new plant lectin family.* Nature structural biology, *2*(6):472–9, 1995. doi:10.1038/nsb0695-472.

Hildebrand A., Remmert M., Biegert A., and Söding J. *Fast and accurate automatic structure prediction with HHpred.* Proteins: Structure, Function and Bioinformatics, *77*:128–132, 2009. doi:10.1002/prot.22499.

Hofmann K. and Bucher P. *The rsp5-domain is shared by proteins of diverse functions.* FEBS letters, *358*(2):153–7, 1995. doi:10.1016/0014-5793(94)01415-W.

Holland T.A., Veretnik S., Shindyalov I.N., and Bourne P.E. *Partitioning protein structures into domains: why is it so difficult?* Journal of molecular biology, *361*(3):562–90, 2006. doi:10.1016/j.jmb.2006.05.060.

Holm L. and Rosenström P. *Dali server: conservation mapping in 3D.* Nucleic acids research, *38*(Web Server issue):W545–9, 2010. doi:10.1093/nar/gkq366.

Hoorelbeke B., Van Damme E.J.M., Rougé P., Schols D., Van Laethem K., Fouquaert E., and Balzarini J. *Differences in the mannose oligomer specificities of the closely related lectins from Galanthus nivalis and Zea mays strongly determine their eventual anti-HIV activity.* Retrovirology, *8*(1):10, 2011. doi:10.1186/1742-4690-8-10.

Hsieh T.J., Farh L., Huang W.M., and Chan N.L. *Structure of the topoisomerase IV C-terminal domain: a broken beta-propeller implies a role as geometry facilitator in catalysis.* The Journal of biological chemistry, *279*(53):55587–93, 2004. doi:10.1074/jbc.M408934200.

Huang J., Gurung B., Wan B., Matkar S., Veniaminova N.A., Wan K., Merchant J.L., Hua X., and Lei M. *The same pocket in menin binds both MLL and JUND but has opposite effects on transcription.* Nature, *482*(7386):542–6, 2012. doi:10.1038/nature10806.

Huang Y.J., Mao B., Aramini J.M., and Montelione G.T. *Assessment of template-based protein structure predictions in CASP10.* Proteins, *82 Suppl 2*:43–56, 2014. doi:10.1002/prot.24488.

Hurley J.H. and Hanson P.I. *Membrane budding and scission by the ESCRT machinery: it's all in the neck.* Nature reviews. Molecular cell biology, *11*(8):556–66, 2010. doi:10.1038/nrm2937.

Iwaya N., Takasu H., Goda N., Shirakawa M., Tanaka T., Hamada D., and Hiroaki H. *MIT domain of Vps4 is a Ca2+-dependent phosphoinositide-binding domain.* Journal of biochemistry, *153*(5):473–81, 2013. doi:10.1093/jb/mvt012.

Jeffares D.C., Poole A.M., and Penny D. *Relics from the RNA world.* Journal of molecular evolution, *46*(1):18–36, 1998. doi:10.1007/PL00006280.

Jones E., Oliphant T., Peterson P., and Others. *SciPy: Open source scientific tools for Python*. 2001.

Joyce G.F. *The antiquity of RNA-based evolution.* Nature, *418*(6894):214–21, 2002. doi:10.1038/418214a.

Kabsch W. *A solution for the best rotation to relate two sets of vectors*. Acta Crystallographica Section A, *32*(5):922–923, 1976. doi:10.1107/S0567739476001873.

Kabsch W. *A discussion of the solution for the best rotation to relate two sets of vectors*. Acta Crystallographica Section A, *34*(5):827–828, 1978. doi:10.1107/S0567739478001680.

Kalev I., Mechelke M., Kopec K.O., Holder T., Carstens S., and Habeck M. *CSB: a Python framework for structural bioinformatics.* Bioinformatics (Oxford, England), *28*(22):2996–7, 2012. doi:10.1093/bioinformatics/bts538.

Kampranis S.C. and Maxwell A. *Conversion of DNA gyrase into a conventional type II topoisomerase.* Proceedings of the National Academy of Sciences of the United States of America, *93*(25):14416–21, 1996.

Karpenahalli M.R. *Exploring Protein Domain Evolution by Designing New TPR-like Domains*. Ph.D. thesis, Max-Planck-Institute for Developmental Biology, 2006.

Karpenahalli M.R., Lupas A.N., and Söding J. *TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences.* BMC bioinformatics, *8*:2, 2007. doi:10.1186/1471-2105-8-2.

Katoh K., Suzuki H., Terasawa Y., Mizuno T., Yasuda J., Shibata H., and Maki M. *The penta-EF-hand protein ALG-2 interacts directly with the ESCRT-I component TSG101, and Ca2+-dependently co-localizes to aberrant endosomes with dominant-negative AAA ATPase SKD1/Vps4B.* The Biochemical journal, *391*(Pt 3):677–85, 2005. doi:10.1042/BJ20050398.

Kelley L.A. and Sternberg M.J.E. *Protein structure prediction on the Web: a case study using the Phyre server.* Nature protocols, *4*(3):363–71, 2009. doi:10.1038/nprot.2009.2.

Kendrew J.C., Bodo G., Dintzis H.M., Parrish R.G., Wyckoff H., and Phillips D.C. *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.* Nature, *181*(4610):662–6, 1958.

Kim J., Sitaraman S., Hierro A., Beach B.M., Odorizzi G., and Hurley J.H. *Structural basis for endosomal targeting by the Bro1 domain.* Developmental cell, *8*(6):937–47, 2005. doi:10.1016/j.devcel.2005.04.001.

Kimura M. *Evolutionary Rate at the Molecular Level*. Nature, *217*(5129):624–626, 1968. doi:10.1038/217624a0.

Kleiger G., Beamer L.J., Grothe R., Mallick P., and Eisenberg D. *The 1.7 A crystal structure of BPI: a study of how two dissimilar amino acid sequences can adopt the same fold.* Journal of molecular biology, *299*(4):1019–34, 2000. doi:10.1006/jmbi.2000.3805.

Koch O. and Klebe G. *Turns revisited: a uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions.* Proteins, *74*(2):353–67, 2009. doi:10.1002/prot.22185.

Kolodziejczyk R., Bujacz G., Jakób M., Ozyhar A., Jaskolski M., and Kochman M. *Insect juvenile hormone binding protein shows ancestral fold present in human lipid-binding proteins.* Journal of molecular biology, *377*(3):870–81, 2008. doi:10.1016/j.jmb.2008.01.026.

Kopec K.O., Alva V., and Lupas A.N. *Homology of SMP domains to the TULIP superfamily of lipid-binding proteins provides a structural basis for lipid exchange between ER and mitochondria.* Bioinformatics (Oxford, England), *26*(16):1927–31, 2010. doi:10.1093/bioinformatics/btq326.

Kopec K.O., Alva V., and Lupas A.N. *Bioinformatics of the TULIP domain superfamily.* Biochemical Society transactions, *39*(4):1033–8, 2011. doi:10.1042/BST0391033.

Kopec K.O. and Lupas A.N. *β-Propeller Blades as Ancestral Peptides in Protein Evolution*. PLoS ONE, *8*(10):e77074, 2013. doi:10.1371/journal.pone.0077074.

Kornmann B., Currie E., Collins S.R., Schuldiner M., Nunnari J., Weissman J.S., and Walter P. *An ER-mitochondria tethering complex revealed by a synthetic biology screen.* Science (New York, N.Y.), *325*(5939):477–81, 2009. doi:10.1126/science.1175088.

Koropatkin N.M., Martens E.C., Gordon J.I., and Smith T.J. *Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices.* Structure (London, England : 1993), *16*(7):1105–15, 2008. doi:10.1016/j.str.2008.03.017.

Kozlov G., Denisov A.Y., Girard M., Dicaire M.J., Hamlin J., McPherson P.S., Brais B., and Gehring K. *Structural basis of defects in the sacsin HEPN domain responsible for autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS).* The Journal of biological chemistry, *286*(23):20407–12, 2011. doi:10.1074/jbc.M111.232884.

Kramlinger V.M. and Hiasa H. *The "GyrA-box" is required for the ability of DNA gyrase to wrap DNA and catalyze the supercoiling reaction.* The Journal of biological chemistry, *281*(6):3738–42, 2006. doi:10.1074/jbc.M511160200.

Krishna S.S. and Grishin N.V. *Structurally analogous proteins do exist!* Structure (London, England : 1993), *12*(7):1125–7, 2004. doi:10.1016/j.str.2004.06.004.

Kryshtafovych A., Krysko O., Daniluk P., Dmytriv Z., and Fidelis K. *Protein structure prediction center in CASP8.* Proteins, *77 Suppl 9*:5–9, 2009. doi:10.1002/prot.22517.

Kurimoto E., Suzuki M., Amemiya E., Yamaguchi Y., Nirasawa S., Shimba N., Xu N., Kashiwagi T., Kawai M., Suzuki E.i., *et al. Curculin exhibits sweet-tasting and taste-modifying activities through its distinct molecular surfaces.* The Journal of biological chemistry, *282*(46):33252–6, 2007. doi:10.1074/jbc.C700174200.

Lam H., Oh D.C., Cava F., Takacs C.N., Clardy J., de Pedro M.A., and Waldor M.K. *D-amino acids govern stationary phase cell wall remodeling in bacteria.* Science (New York, N.Y.), *325*(5947):1552–5, 2009. doi:10.1126/science.1178123.

Lee I. and Hong W. *Diverse membrane-associated proteins contain a novel SMP domain.* FASEB journal : official publication of the Federation of American Societies for Experimental Biology, *20*(2):202–6, 2006. doi:10.1096/fj.05-4581hyp.

Legrand P., Pinaud N., Minvielle-Sébastia L., and Fribourg S. *The structure of the CstF-77 homodimer provides insights into CstF assembly.* Nucleic acids research, *35*(13):4515–22, 2007. doi:10.1093/nar/gkm458.

Levinthal C. *How to fold graciously*. In *Mössbaun Spectroscopy in Biological Systems Proceedings*, volume 24, pages 22–24. 1969.

Li M.H., Lin L., Wang X.L., and Liu T. *Protein-protein interaction site prediction based on conditional random fields.* Bioinformatics (Oxford, England), *23*(5):597–604, 2007. doi:10.1093/bioinformatics/btl660.

Li Y. and Romeis J. *Impact of snowdrop lectin (Galanthus nivalis agglutinin; GNA) on adults of the green lacewing, Chrysoperla carnea.* Journal of insect physiology, *55*(2):135–42, 2009. doi:10.1016/j.jinsphys.2008.10.015.

Lill R. and Kispal G. *Maturation of cellular Fe-S proteins: an essential function of mitochondria.* Trends in biochemical sciences, *25*(8):352–6, 2000. doi:10.1016/S0968-0004(00)01589-9.

Linderstrøm-Lang K.U. *Proteins and Enzymes*, volume 6 of *Lane Medical Lectures, Stanford University Publications, University Series, Medical Sciences*. Stanford University Press, Stanford, CA, 1952.

Liu S., Rauhut R., Vornlocher H.P., and Lührmann R. *The network of protein-protein interactions within the human U4/U6.U5 tri-snRNP.* RNA (New York, N.Y.), *12*(7):1418–30, 2006. doi:10.1261/rna.55406.

Liu W., Yang N., Ding J., Huang R.h., Hu Z., and Wang D.C. *Structural mechanism governing the quaternary organization of monocot mannose-binding lectin revealed by the novel monomeric structure of an orchid lectin.* The Journal of biological chemistry, *280*(15):14865–76, 2005. doi:10.1074/jbc.M411634200.

Lupas A.N. and Koretke K.K. *Evolution of Protein Folds*. In T. Schwede and M.C. Peitsch, editors, *Computational Structural Biology: Methods and Applications*, chapter Section II, pages 131–151. World Scientific Publishing Company, Incorporated, Singapore, 2008. ISBN 978-981-277-877-2. doi:10.1142/9789812778789_0006.

Lupas A.N., Ponting C.P., and Russell R.B. *On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?* Journal of structural biology, *134*(2-3):191–203, 2001. doi:10.1006/jsbi.2001.4393.

Lupyan D., Leo-Macias A., and Ortiz A.R. *A new progressive-iterative algorithm for multiple structure alignment.* Bioinformatics (Oxford, England), *21*(15):3255–63, 2005. doi:10.1093/bioinformatics/bti527.

Lurin C., Andrés C., Aubourg S., Bellaoui M., Bitton F., Bruyère C., Caboche M., Debast C., Gualberto J., Hoffmann B., *et al. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis.* The Plant cell, *16*(8):2089–103, 2004. doi:10.1105/tpc.104.022236.

Lynch N.R., Thomas W.R., Garcia N.M., Di Prisco M.C., Puccio F.A., L'opez R.I., Hazell L.A., Shen H.D., Lin K.L., and Chua K.Y. *Biological activity of recombinant Der p 2, Der p 5 and Der p 7 allergens of the house-dust mite Dermatophagoides pteronyssinus.* International archives of allergy and immunology, *114*(1):59–67, 1997. doi:10.1159/000237644.

Main E.R.G., Jackson S.E., and Regan L. *The folding and design of repeat proteins: reaching a consensus.* Current opinion in structural biology, *13*(4):482–9, 2003a. doi:10.1016/S0959-440X(03)00105-2.

Main E.R.G., Lowe A.R., Mochrie S.G.J., Jackson S.E., and Regan L. *A recurring theme in protein engineering: the design, stability and folding of repeat proteins.* Current opinion in structural biology, *15*(4):464–71, 2005. doi:10.1016/j.sbi.2005.07.003.

Main E.R.G., Xiong Y., Cocco M.J., D'Andrea L.D., and Regan L. *Design of stable alpha-helical arrays from an idealized TPR motif.* Structure (London, England : 1993), *11*(5):497–508, 2003b. doi:10.1016/S0969-2126(03)00076-5.

Mariani V., Kiefer F., Schmidt T., Haas J., and Schwede T. *Assessment of template based protein structure predictions in CASP9.* Proteins, *79 Suppl 1*:37–58, 2011. doi:10.1002/prot.23177.

Martens E.C., Koropatkin N.M., Smith T.J., and Gordon J.I. *Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm.* The Journal of biological chemistry, *284*(37):24673–7, 2009. doi:10.1074/jbc.R109.022848.

Mashaghi A., Kramer G., Lamb D.C., Mayer M.P., and Tans S.J. *Chaperone action at the single-molecule level.* Chemical reviews, *114*(1):660–76, 2014. doi:10.1021/cr400326k.

McBride H.M., Neuspiel M., and Wasiak S. *Mitochondria: more than just a powerhouse.* Current biology : CB, *16*(14):R551–60, 2006. doi:10.1016/j.cub.2006.06.054.

Menke M., Berger B., and Cowen L.J. *Markov random fields reveal an N-terminal double beta-propeller motif as part of a bacterial hybrid two-component sensor system.* Proceedings of the National Academy of Sciences of the United States of America, *107*(9):4069–74, 2010. doi:10.1073/pnas.0909950107.

Minor D.L. and Kim P.S. *Measurement of the beta-sheet-forming propensities of amino acids.* Nature, *367*(6464):660–3, 1994. doi:10.1038/367660a0.

Moore A.D., Björklund A.K., Ekman D., Bornberg-Bauer E., and Elofsson A. *Arrangements in the modular evolution of proteins.* Trends in biochemical sciences, *33*(9):444–51, 2008. doi:10.1016/j.tibs.2008.05.008.

Mueller G.A., Edwards L.L., Aloor J.J., Fessler M.B., Glesner J., Pomés A., Chapman M.D., London R.E., and Pedersen L.C. *The structure of the dust mite allergen Der p 7 reveals similarities to innate immune proteins.* The Journal of allergy and clinical immunology, *125*(4):909–917.e4, 2010. doi:10.1016/j.jaci.2009.12.016.

Murzin A.G., Brenner S.E., Hubbard T.J.P., and Chothia C. *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* Journal of molecular biology, *247*(4):536–40, 1995. doi:10.1006/jmbi.1995.0159.

Needleman S.B. and Wunsch C.D. *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* Journal of molecular biology, *48*(3):443–53, 1970. doi:10.1016/0022-2836(70)90057-4.

Obsil T., Ghirlando R., Klein D.C., Ganguly S., and Dyda F. *Crystal Structure of the 14-3-3ζ:Serotonin N-Acetyltransferase Complex*. Cell, *105*(2):257–267, 2001. doi:10.1016/S0092-8674(01)00316-6.

Obsil T. and Obsilova V. *Structural basis of 14-3-3 protein functions.* Seminars in cell & developmental biology, *22*(7):663–72, 2011. doi:10.1016/j.semcdb.2011.09.001.

Opitz C.A., Kulke M., Leake M.C., Neagoe C., Hinssen H., Hajjar R.J., and Linke W.A. *Damped elastic recoil of the titin spring in myofibrils of human myocardium.* Proceedings of the National Academy of Sciences of the United States of America, *100*(22):12688–93, 2003. doi:10.1073/pnas.2133733100.

Orgel L.E. *Prebiotic chemistry and the origin of the RNA world.* Critical reviews in biochemistry and molecular biology, *39*(2):99–123, 2004. doi:10.1080/10409230490460765.

Pace C.N. and Scholtz J.M. *A helix propensity scale based on experimental studies of peptides and proteins.* Biophysical journal, *75*(1):422–7, 1998. doi:10.1016/S0006-3495(98)77529-0.

Paoli M., Liddington R., Tame J., Wilkinson A., and Dodson G. *Crystal structure of T state haemoglobin with oxygen bound at all four haems.* Journal of molecular biology, *256*(4):775–92, 1996. doi:10.1006/jmbi.1996.0124.

Pasek S., Risler J.L., and Brézellec P. *Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins.* Bioinformatics (Oxford, England), *22*(12):1418–23, 2006. doi:10.1093/bioinformatics/btl135.

Pathare G.R., Nagy I., Bohn S., Unverdorben P., Hubert A., Körner R., Nickell S., Lasker K., Sali A., Tamura T., *et al. The proteasomal subunit Rpn6 is a molecular clamp holding the core and regulatory subcomplexes together.* Proceedings of the National Academy of Sciences of the United States of America, *109*(1):149–54, 2012. doi:10.1073/pnas.1117648108.

Pauling L.C., Corey R.B., and Branson H.R. *The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain.* Proceedings of the National Academy of Sciences of the United States of America, *37*(4):205–11, 1951. doi:10.1073/pnas.37.4.205.

Pearson W.R. and Lipman D.J. *Improved tools for biological sequence comparison.* Proceedings of the National Academy of Sciences of the United States of America, *85*(8):2444–8, 1988.

Pei J. and Grishin N.V. *AL2CO: calculation of positional conservation in a protein sequence alignment.* Bioinformatics (Oxford, England), *17*(8):700–12, 2001. doi:10.1093/bioinformatics/17.8.700.

Perry A.J., Hulett J.M., Likić V.A., Lithgow T., and Gooley P.R. *Convergent evolution of receptors for protein import into mitochondria.* Current biology : CB, *16*(3):221–9, 2006. doi:10.1016/j.cub.2005.12.034.

Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., and Ferrin T.E. *UCSF Chimera–a visualization system for exploratory research and analysis.* Journal of computational chemistry, *25*:1605–1612, 2004. doi:10.1002/jcc.20084.

Phillips S.A., Barr V.A., Haft D.H., Taylor S.I., and Haft C.R. *Identification and characterization of SNX15, a novel sorting nexin involved in protein trafficking.* The Journal of biological chemistry, *276*(7):5074–84, 2001. doi:10.1074/jbc.M004671200.

Pisarewicz K., Mora D., Pflueger F.C., Fields G.B., and Marí F. *Polypeptide chains containing D-gamma-hydroxyvaline.* Journal of the American Chemical Society, *127*(17):6207–15, 2005. doi:10.1021/ja050088m.

Ponting C.P. *Proteins of the endoplasmic-reticulum-associated degradation pathway: domain detection and function prediction.* The Biochemical journal, *351 Pt 2*(Pt 2):527–35, 2000.

Preker P.J. and Keller W. *The HAT helix, a repetitive motif implicated in RNA processing.* Trends in biochemical sciences, *23*(1):15–6, 1998. doi:10.1016/S0968-0004(97)01156-0.

Qi Y., Pei J., and Grishin N.V. *C-terminal domain of gyrase A is predicted to have a beta-propeller structure.* Proteins, *47*(3):258–64, 2002. doi:10.1002/prot.10090.

Qiu X., Mistry A., Ammirati M.J., Chrunyk B.A., Clark R.W., Cong Y., Culp J.S., Danley D.E., Freeman T.B., Geoghegan K.F., *et al. Crystal structure of cholesteryl ester transfer protein reveals a long tunnel and four bound lipid molecules.* Nature structural & molecular biology, *14*(2):106–13, 2007. doi:10.1038/nsmb1197.

Quistgaard E.M., Madsen P., Grøftehauge M.K., Nissen P., Petersen C.M., and Thirup S.S. *Ligands bind to Sortilin in the tunnel of a ten-bladed beta-propeller domain.* Nature structural & molecular biology, *16*(1):96–8, 2009. doi:10.1038/nsmb.1543.

Ramachandraiah G. and Chandra N. *Sequence and structural determinants of mannose recognition.* Proteins, *39*(4):358–64, 2000.

Remmert M., Biegert A., Hauser A., and Söding J. *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nature methods, *9*(2):173–5, 2012. doi:10.1038/nmeth.1818.

Remmert M., Biegert A., Linke D., Lupas A.N., and Söding J. *Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin.* Molecular biology and evolution, *27*(6):1348–58, 2010. doi:10.1093/molbev/msq017.

Richardson J.S. and Richardson D.C. *Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation.* Proceedings of the National Academy of Sciences of the United States of America, *99*(5):2754–9, 2002. doi:10.1073/pnas.052706099.

Robillard G.T. and Broos J. *Structure/function studies on the bacterial carbohydrate transporters, enzymes II, of the phosphoenolpyruvate-dependent phosphotransferase system.* Biochimica et biophysica acta, *1422*(2):73–104, 1999. doi:10.1016/S0304-4157(99)00002-7.

Ron D. and Walter P. *Signal integration in the endoplasmic reticulum unfolded protein response.* Nature reviews. Molecular cell biology, *8*(7):519–29, 2007. doi:10.1038/nrm2199.

Roy A., Kucukural A., and Zhang Y. *I-TASSER: a unified platform for automated protein structure and function prediction.* Nature protocols, *5*(4):725–38, 2010. doi:10.1038/nprot.2010.5.

Russell R.B., Saqi M.A., Sayle R.A., Bates P.A., and Sternberg M.J.E. *Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation.* Journal of molecular biology, *269*(3):423–39, 1997. doi:10.1006/jmbi.1997.1019.

Sadreyev R. and Grishin N.V. *COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance.* Journal of molecular biology, *326*(1):317–36, 2003. doi:10.1016/S0022-2836(02)01371-2.

Salem G.M., Hutchinson E.G., Orengo C.A., and Thornton J.M. *Correlation of observed fold frequency with the occurrence of local structural motifs.* Journal of molecular biology, *287*(5):969–81, 1999. doi:10.1006/jmbi.1999.2642.

Sarov-Blat L., So W.V., Liu L., and Rosbash M. *The Drosophila takeout gene is a novel molecular link between circadian rhythms and feeding behavior.* Cell, *101*(6):647–56, 2000. doi:10.1016/S0092-8674(00)80876-4.

Schauder C.M., Wu X., Saheki Y., Narayanaswamy P., Torta F., Wenk M.R., De Camilli P., and Reinisch K.M. *Structure of a lipid-bound extended synaptotagmin indicates a role in lipid transfer.* Nature, *510*(7506):552–5, 2014. doi:10.1038/nature13269.

Scheufler C., Brinker A., Bourenkov G., Pegoraro S., Moroder L., Bartunik H., Hartl F.U., and Moarefi I. *Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine.* Cell, *101*(2):199–210, 2000. doi:10.1016/S0092-8674(00)80830-2.

Schoeffler A.J. and Berger J.M. *DNA topoisomerases: harnessing and constraining energy to govern chromosome topology.* Quarterly reviews of biophysics, *41*(1):41–101, 2008. doi:10.1017/S003358350800468X.

Scott A., Chung H.Y., Gonciarz-Swiatek M., Hill G.C., Whitby F.G., Gaspar J., Holton J.M., Viswanathan R., Ghaffarian S., Hill C.P., *et al. Structural and mechanistic studies of VPS4 proteins.* The EMBO journal, *24*(20):3658–69, 2005a. doi:10.1038/sj.emboj.7600818.

Scott A., Gaspar J., Stuchell-Brereton M.D., Alam S.L., Skalicky J.J., and Sundquist W.I. *Structure and ESCRT-III protein interactions of the MIT domain of human VPS4A.* Proceedings of the National Academy of Sciences of the United States of America, *102*(39):13813–8, 2005b. doi:10.1073/pnas.0502165102.

Selkoe D.J. *Cell biology of protein misfolding: the examples of Alzheimer's and Parkinson's diseases.* Nature cell biology, *6*(11):1054–61, 2004. doi:10.1038/ncb1104-1054.

Sharma A., Chandran D., Singh D.D., and Vijayan M. *Multiplicity of carbohydrate-binding sites in β-prism fold lectins: occurrence and possible evolutionary implications*. Journal of Biosciences, *32*(S2):1089–1110, 2008. doi:10.1007/s12038-007-0111-3.

Shen H.D., Chua K.Y., Lin W.L., Chen H.L., Hsieh K.H., and Thomas W.R. *IgE and monoclonal antibody binding by the mite allergen Der p 7.* Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology, *26*(3):308–15, 1996. doi:10.1111/j.1365-2222.1996.tb00096.x.

Shen H.D., Chua K.Y., Lin W.L., Hsieh K.H., and Thomas W.R. *Characterization of the house dust mite allergen Der p 7 by monoclonal antibodies.* Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology, *25*(5):416–22, 1995.

Shetty K.N., Bhat G.G., Inamdar S.R., Swamy B.M., and Suguna K. *Crystal structure of a β-prism II lectin from Remusatia vivipara.* Glycobiology, *22*(1):56–69, 2012. doi:10.1093/glycob/cwr100.

Shieh A.D., Hashimoto T.B., and Airoldi E.M. *Tree preserving embedding.* Proceedings of the National Academy of Sciences of the United States of America, *108*(41):16916–21, 2011. doi:10.1073/pnas.1018393108.

Sikorski R.S., Boguski M.S., Goebl M., and Hieter P. *A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis.* Cell, *60*(2):307–17, 1990. doi:10.1016/0092-8674(90)90745-Z.

Sillitoe I., Cuff A.L., Dessailly B.H., Dawson N.L., Furnham N., Lee D., Lees J.G., Lewis T.E., Studer R.A., Rentzsch R., *et al. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures.* Nucleic acids research, *41*(Database issue):D490–8, 2013. doi:10.1093/nar/gks1211.

Sliz P., Engelmann R., Hengstenberg W., and Pai E.F. *The structure of enzyme IIAlactose from Lactococcus lactis reveals a new fold and points to possible interactions of a multicomponent system.* Structure (London, England : 1993), *5*(6):775–88, 1997.

Small I. *The PPR motif - a TPR-related motif prevalent in plant organellar proteins*. Trends in Biochemical Sciences, *25*(2):45–47, 2000. doi:10.1016/S0968-0004(99)01520-0.

Smith T.F. and Waterman M.S. *Identification of common molecular subsequences.* Journal of molecular biology, *147*(1):195–7, 1981. doi:10.1016/0022-2836(81)90087-5.

Söding J. *Protein homology detection by HMM-HMM comparison.* Bioinformatics (Oxford, England), *21*(7):951–60, 2005. doi:10.1093/bioinformatics/bti125.

Söding J., Biegert A., and Lupas A.N. *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic acids research, *33*(Web Server issue):W244–8, 2005. doi:10.1093/nar/gki408.

Söding J. and Lupas A.N. *More than the sum of their parts: on the evolution of proteins from peptides.* BioEssays : news and reviews in molecular, cellular and developmental biology, *25*(9):837–46, 2003. doi:10.1002/bies.10321.

Söding J., Remmert M., Biegert A., and Lupas A.N. *HHsenser: exhaustive transitive profile search using HMM-HMM comparison.* Nucleic acids research, *34*(Web Server issue):W374–8, 2006. doi:10.1093/nar/gkl195.

Stevens T.J. and Paoli M. *RCC1-like repeat proteins: a pangenomic, structurally diverse new superfamily of beta-propeller domains.* Proteins, *70*(2):378–87, 2008. doi:10.1002/prot.21521.

Sudol M., Chen H.I., Bougeret C., Einbond A., and Bork P. *Characterization of a novel protein-binding module–the WW domain.* FEBS letters, *369*(1):67–71, 1995. doi:10.1016/0014-5793(95)00550-S.

Sudol M., Recinos C.C., Abraczinskas J., Humbert J., and Farooq A. *WW or WoW: the WW domains in a union of bliss.* IUBMB life, *57*(12):773–8, 2005. doi:10.1080/15216540500389039.

Suzuki R., Fujimoto Z., Shiotsuki T., Tsuchiya W., Momma M., Tase A., Miyazawa M., and Yamazaki T. *Structural mechanism of JH delivery in hemolymph by JHBP of silkworm, Bombyx mori.* Scientific reports, *1*:133, 2011. doi:10.1038/srep00133.

Szklarczyk R. and Heringa J. *Tracking repeats using significance and transitivity.* Bioinformatics (Oxford, England), *20 Suppl 1*:i311–7, 2004. doi:10.1093/bioinformatics/bth911.

Tang C., Williams D.C., Ghirlando R., and Clore G.M. *Solution structure of enzyme IIA(Chitobiose) from the N,N'-diacetylchitobiose branch of the Escherichia coli phosphotransferase system.* The Journal of biological chemistry, *280*(12):11770–80, 2005. doi:10.1074/jbc.M414300200.

ter Haar E., Musacchio A., Harrison S.C., and Kirchhausen T. *Atomic structure of clathrin: a beta propeller terminal domain joins an alpha zigzag linker.* Cell, *95*(4):563–73, 1998. doi:10.1016/S0092-8674(00)81623-2.

Thakker R.V. *Multiple endocrine neoplasia type 1 (MEN1).* Best practice & research. Clinical endocrinology & metabolism, *24*(3):355–70, 2010. doi:10.1016/j.beem.2010.07.003.

Thomas W.R. and Hales B.J. *T and B cell responses to HDM allergens and antigens.* Immunologic research, *37*(3):187–99, 2007.

Tzivion G. and Avruch J. *14-3-3 proteins: active cofactors in cellular regulation by serine/threonine phosphorylation.* The Journal of biological chemistry, *277*(5):3061–4, 2002. doi:10.1074/jbc.R100059200.

Unverdorben P., Beck F., Śledź P., Schweitzer A., Pfeifer G., Plitzko J.M., Baumeister W., and Förster F. *Deep classification of a large cryo-EM dataset defines the conformational landscape of the 26S proteasome.* Proceedings of the National Academy of Sciences of the United States of America, *111*(15):5544–9, 2014. doi:10.1073/pnas.1403409111.

van der Maaten L. and Hinton G. *Visualizing Data using t-SNE*. Journal of Machine Learning Research, *9*:2579–2605, 2008.

Venkatachalam C.M. *Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units.* Biopolymers, *6*(10):1425–36, 1968. doi:10.1002/bip.1968.360061006.

Viterbi A.J. *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* IEEE Transactions on Information Theory, *13*(2):260–269, 1967. doi:10.1109/TIT.1967.1054010.

Voelker D.R. *New perspectives on the regulation of intermembrane glycerophospholipid traffic.* Journal of lipid research, *44*(3):441–9, 2003. doi:10.1194/jlr.R200020-JLR200.

Wang Z., Eickholt J., and Cheng J. *MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8.* Bioinformatics (Oxford, England), *26*(7):882–8, 2010. doi:10.1093/bioinformatics/btq058.

Webb C., Upadhyay A., Giuntini F., Eggleston I., Furutani-Seiki M., Ishima R., and Bagby S. *Structural features and ligand binding properties of tandem WW domains from YAP and TAZ, nuclear effectors of the Hippo pathway.* Biochemistry, *50*(16):3300–9, 2011. doi:10.1021/bi2001888.

Wolosker H., Dumin E., Balan L., and Foltyn V.N. *D-amino acids in the brain: D-serine in neurotransmission and neurodegeneration.* The FEBS journal, *275*(14):3514–26, 2008. doi:10.1111/j.1742-4658.2008.06515.x.

Wu S. and Zhang Y. *MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information.* Proteins, *72*(2):547–56, 2008. doi:10.1002/prot.21945.

Xu J. and Zhang Y. *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics (Oxford, England), *26*(7):889–95, 2010. doi:10.1093/bioinformatics/btq066.

Yagi Y., Hayashi S., Kobayashi K., Hirayama T., and Nakamura T. *Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants.* PloS ONE, *8*(3):e57286, 2013. doi:10.1371/journal.pone.0057286.

Yang Y. and Hua X. *In search of tumor suppressing functions of menin.* Molecular and cellular endocrinology, *265-266*:34–41, 2007. doi:10.1016/j.mce.2006.12.032.

Ybe J.A., Brodsky F.M., Hofmann K., Lin K., Liu S.H., Chen L., Earnest T.N., Fletterick R.J., and Hwang P.K. *Clathrin self-assembly is mediated by a tandemly repeated superhelix.* Nature, *399*(6734):371–5, 1999. doi:10.1038/20708.

Yin P., Li Q., Yan C., Liu Y., Liu J., Yu F., Wang Z., Long J., He J., Wang H.W., *et al. Structural basis for the modular recognition of single-stranded RNA by PPR proteins.* Nature, *504*(7478):168–71, 2013. doi:10.1038/nature12651.

Yona G. and Levitt M. *Within the twilight zone: a sensitive profile-profile comparison tool based on information theory.* Journal of molecular biology, *315*(5):1257–75, 2002. doi:10.1006/jmbi.2001.5293.

Youngman M.J., Hobbs A.E.A., Burgess S.M., Srinivasan M., and Jensen R.E. *Mmm2p, a mitochondrial outer membrane protein required for yeast mitochondrial shape and maintenance of mtDNA nucleoids.* The Journal of cell biology, *164*(5):677–88, 2004. doi:10.1083/jcb.200308012.

Zhang H. and Grishin N.V. *The alpha-subunit of protein prenyltransferases is a member of the tetratricopeptide repeat family.* Protein science : a publication of the Protein Society, *8*(8):1658–67, 1999. doi:10.1110/ps.8.8.1658.

Zhang Y. *Template-based modeling and free modeling by I-TASSER in CASP7.* Proteins, *69 Suppl 8*:108–17, 2007. doi:10.1002/prot.21702.

Zhang Y. *I-TASSER: fully automated protein structure prediction in CASP8.* Proteins, *77 Suppl 9*:100–13, 2009. doi:10.1002/prot.22588.

Zhang Y. and Skolnick J. *Scoring function for automated assessment of protein structure template quality.* Proteins, *57*(4):702–10, 2004. doi:10.1002/prot.20264.

Zhang Y. and Skolnick J. *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic acids research, *33*(7):2302–9, 2005. doi:10.1093/nar/gki524.

Zhou J., Liu C.Y., Back S.H., Clark R.L., Peisach D., Xu Z., and Kaufman R.J. *The crystal structure of human IRE1 luminal domain reveals a conserved dimerization interface required for activation of the unfolded protein response.* Proceedings of the National Academy of Sciences of the United States of America, *103*(39):14343–8, 2006. doi:10.1073/pnas.0606480103.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`http://code.google.com/p/classicthesis/`