memorec biotec gmbh
a miltenyi biotec company

Comparative Analysis of the Ubiquitin-proteasome system in *Homo sapiens* and *Saccharomyces cerevisiae*

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Hartmut Scheel
aus Rheinbach

Köln, 2005

Berichterstatter:

Prof. Dr. R. Jürgen Dohmen

Prof. Dr. Thomas Langer

Dr. Kay Hofmann

Tag der mündlichen Prüfung: 18.07.2005

Natura non saltat
(Linné)

Zusammenfassung

Das Ubiquitin-Proteasom System (UPS) stellt den wichtigsten Abbauweg für intrazelluläre Proteine in eukaryotischen Zellen dar. Das abzubauende Protein wird zunächst über eine Enzym-Kaskade mit einer kovalent gebundenen Ubiquitinkette markiert. Anschließend wird das konjugierte Substrat vom Proteasom erkannt und proteolytisch gespalten. Ubiquitin besitzt eine Reihe von Homologen, die ebenfalls posttranslational an Proteine gekoppelt werden können, wie z.B. SUMO und NEDD8. Die hierbei verwendeten Aktivierungs- und Konjugations-Kaskaden sind vollständig analog zu der des Ubiquitin-Systems. Es ist charakteristisch für das UPS, daß sich die Vielzahl der daran beteiligten Proteine aus nur wenigen Proteinfamilien rekrutiert, die durch gemeinsame, funktionale Homologiedomänen gekennzeichnet sind. Einige dieser funktionalen Domänen sind auch in den Modifikations-Systemen der Ubiquitin-Homologen zu finden, jedoch verfügen diese Systeme zusätzlich über spezifische Domänentypen.

Homologiedomänen lassen sich als mathematische Modelle in Form von Domänen-deskriptoren (Profile) beschreiben. Diese Deskriptoren können wiederum dazu verwendet werden, mit Hilfe geeigneter Verfahren eine gegebene Proteinsequenz auf das Vorliegen von entsprechenden Homologiedomänen zu untersuchen. Da die im UPS involvierten Homologie-domänen fast ausschließlich auf dieses System und seine Analoga beschränkt sind, können domänen-spezifische Profile zur Katalogisierung der UPS-relevanten Proteine einer Spezies verwendet werden. Auf dieser Basis können dann die entsprechenden UPS-Repertoires verschiedener Spezies miteinander verglichen werden.

In dieser Arbeit wurden basierend auf UPS-relevanten Homologiedomänen und unter Verwendung der Profilmethode solche Kataloge für den Menschen und die Hefe *Saccharomyces cerevisiae* erstellt. In Kombination mit phylogenetischen Methoden wurden die evolutionären Beziehungen zwischen den UPS-Komponenten dieser beiden Organismen untersucht und in geeigneten Fällen eine Orthologiebeziehung abgeleitet. Durch die Verwendung der hoch-sensitiven Profiltechnik und die Einbeziehung von genomischen Datenbanken wurden im Rahmen dieser Arbeit eine Reihe von Proteinen identifiziert, die bisher nicht mit dem UPS assoziiert worden waren. Zusätzlich konnten einige unerwartete Verwandtschaftsbeziehungen zwischen Proteinen des UPS abgeleitet werden. So konnte z.B. das lange gesuchte Hefe-Ortholog des 'Antizyms' der Ornithin- Decarboxylase identifiziert werden - eine wichtige Voraussetzung zur experimentellen Untersuchung des Ubiquitin-

unabhängigen Proteinabbaus durch das Proteasom. In einem weiteren Beispiel konnte gezeigt werden, daß Ataxin-3 aus Mensch eine Homologiedomäne mit funktioneller Ähnlichkeit zu den deubiquitylierenden Enzymen besitzt. Da Ataxin-3 bei Patienten mit spinocerebellarer Ataxie 3 (SCA3) mutiert ist, kann diese Entdeckung zur Aufklärung des Krankheitsmechanismus von SCA3 beitragen. In einer dritten exemplarischen Anwendung konnten weitreichende Vorhersagen für den strukturellen Aufbau des 19S-Proteasoms getroffen werden, insbesondere mit Bezug auf dessen 'lid' Subkomplex.

Ein Vergleich der UPS-relevanten Proteinrepertoires der Hefe und des Menschen erlaubte Schlüsse über den evolutionären Ursprung einiger Komponenten des UPS. Insbesondere bei Proteinfamilien mit einer etablierten oder angenommenen Rolle in der Substraterkennung und -ubiquitylierung oder im reversen Prozess der Deubiquitylierung findet man beim Menschen eine starke Diversifizierung der Proteinfamilien, während die elementaren Funktionen des UPS durch annähernd vergleichbare Proteinsets ausgeführt werden. Trotz der teilweise erheblich größeren Proteinfamilien im Menschen, konnten nicht allen UPS-assoziierten Proteinen der Hefe humane Orthologe zugeordnet werden, was auf spezifische Prozesse innerhalb des UPS von *S. cerevisiae* hindeutet. Ingesamt überwiegen jedoch die Ähnlichkeiten der beiden Systeme und unterstreichen die Rolle von *S. cerevisiae* als Modellorganismus zur Aufklärung des UPS.

Abstract

The UPS (ubiquitin-proteasome system) is the most important degradation pathway for intracellular proteins in the eukaryotic cell. In a first step, the protein to degrade (substrate) is tagged covalently with a Ubiquitin chain via an enzyme cascade. Subsequently, the Ubiquitin chain is recognized by the proteasome and the substrate is proteolytically cleaved. Ubiquitin has several homologues, which can be conjugated to proteins posttranslationally, e.g. SUMO or NEDD8. The enzymes used for activation and conjugation of the Ubiquitin homologues are completely analogous to the ones used in the UPS. A hallmark of the UPS is that most proteins involved belong to only a few protein families, which are characterized by common functional homology domains. Several of these homology domains are found in the modification systems of Ubiquitin homologues, but these systems appear to have specific homology domains on their own as well.

Homology domains may be described as mathematical models in terms of domain descriptors (profile). These profiles together with appropriate search algorithms can be applied to screen a given protein sequence for the occurrence of the corresponding homology domains. As the homology domains involved in the UPS are almost exclusively found in proteins of this and analogous systems, profiles corresponding to these homology domains seem to be an appropriate means to catalogue proteins of the UPS of a given species.

In this work catalogues of proteins with a known or putative role in the UPS or analogous systems were set up for human and *Saccharomyces cerevisiae* (here referred to as 'yeast') based on relevant homology domains and their corresponding profiles. In combination with phylogenetic methods the evolutionary relationships between the UPS components of these two organisms were analyzed and, if possible, orthologous relationships were derived. Using the highly sensitive profile technique and including genomic databases, several new proteins were identified that have not been associated with the UPS so far. Additionally, several unexpected relationships were revealed between proteins of the UPS. For example, the postulated yeast orthologue of the antizyme of the ornithine decarboxylase could be revealed, which may be important for the experimental analysis of Ubiquitin independent protein degradation by the proteasome. Another example is human ataxin-3, in which a homology domain was found with similarity to the catalytic site of deubiquitylating enzymes. As ataxin-3 is mutated in patients with a spinocerebellar ataxia 3 (SCA3), this discovery might have implications for the elucidation of the SCA3 disease mechanism. Furthermore, predictions on the structure of the 'lid' of the 19S regulatory particle could be formulated.

A comparison of the UPS-relevant protein repertoires of yeast and human allowed conclusions on the evolutionary origin of UPS components. Especially protein families with an established or putative role in substrate recognition/ubiquitylation or in the reverse process of deubiquitylation exhibited a strong diversification in human. Simultaneously, elementary functions of the UPS are carried out by almost identical protein sets in both yeast and human. Despite the extensively expanded protein families in human, not all yeast proteins associated with the UPS could be assigned to human orthologues. This finding might indicate specific processes within the yeast UPS. To summarize, the similarities of both yeast and human UPS are significant and underline the role of *S. Cerevisiae* as a model organism used in analyzing the UPS.

Abbreviations

AAA	ATPases associated with a variety of cellular activities	MPN	Mpr1, Pad1 N-terminal
APC	anaphase-promoting complex (cyclosome)	MSA	multiple sequence alignment
BAG	Bcl2-associated athanogene domain	NEDD	neural precursor expressed, developmentally downregulated
BIRC6	baculoviral IAP repeat-containing protein 6	OAZ	ODC antizyme
BLOSUM	blocks substitution matrix	ODC	ornithine decarboxylase
bp	base pair	ORF	open reading frame
BTB	Bric-a-brac (bab), Tramtrack (ttk), and Broad-Complex (BR-C)	PAM	per cent accepted mutation
cDNA	complementary DNA	PAZ	poly-Ub associated Zn-finger
CHIP	carboxy terminus of hsp70-interacting protein	PCI	proteasome, COP9, initiation factor 3
Clp	caseinolytic protease	PCNA	proliferating cell nuclear antigen
CP	core particle (20S proteasome)	PIAS	protein inhibitor of activated STAT
CSN	COP9 signalosome	POMP	proteasomal maturation protein
CUE	coupling of Ub conjugation to ER degradation	RING	really interesting new gene
Cvt	Cytoplasm-to-vacuole targeting	RP	regulatory particle (19S proteasome subcomplex)
DNA	desoxyribonucleic acid	RPN	regulatory particle non-ATPase
DUB	deubiquitylating enzyme	RPT	regulatory particle triple A ATPase
E1	Ub activating enzyme	SC, sc	Saccharomyces cerevisiae
E2	Ub conjugating enzyme	SCF	Skp1, cullin, F-box (E3 complex)
E3	Ub ligating enzyme	SGD	Saccharomyces Genome Database
E4	Ub chain elongation factor	SUMO	small Ub-like modifier
eIF3	eukaryotic translation initiation factor 3	TrEMBL	translated EMBL nucleotide sequence data library
ENTH	Epsin N-terminal homology domain	Ub	Ubiquitin
ER	endoplasmatic reticulum	UBA	Ubiquitin-pathway associated domain
ERAD	ER-associated protein degradation	UBC	Ub conjugating
EST	expressed sequence tag	UBL	Ub-like modifier
GAT	GGA and TOM (target of myb)	UBP/USP	Ub-specific protease
HAUSP	herpes-associated ubiquitin-specific protease	UBX	Ub-like motif, sometimes referred to as UX domain
HECT	Homologous to the E6-AP Carboxyl Terminus	UCH	Ub C-terminal hydrolase
HMM	hidden Markov model	UEV	Ub conjugating enyzme variant
HS, hs	*Homo sapiens*	UFD	Ub fusion degradation pathway
JAMM	JAB1/MPN/Mov34 metalloenzyme	UIM	Ub interacting motif
kDa	kilodalton	UPS	Ub proteasome system

Contents

1 Introduction

1.1 Comparative sequence analysis

1.1.1 Functional classification of protein sequences

Comparative sequence analysis of proteomes from distinct species is a generally applied approach to derive knowledge on phylogenetic relationships, evolution and function of any new protein sequence. Transfer of available functional information from already characterized proteins to novel ones is often performed based on sequence homology. Homologous protein sequences are sequences that share a common evolutionary ancestor. Homology is often inferred from sequence similarity measurements, although in a few cases sequence similarity seems to have arisen by convergence. Sequences of homologous proteins can diverge greatly over evolutionary time, but function or structure may be maintained anyway. Thus, if sufficient sequence similarity is detected between a well studied and an uncharacterised protein, available information can be transferred along the homologous relationship between the two proteins.

Homologues can be divided into orthologues and paralogues. Orthologues have diverged from each other by a speciation event, i.e. the evolution of new biological species from a common ancestor, while paralogues have diverged from each other by gene duplication events (Fitch, 2000). Unlike orthologues and paralogues, a xenologue represents a homologue that has entered the genome of a species by interspecies gene transfer (horizontal gene transfer). While paralogues often evolve new functions, even if related to the original one, orthologues typically occupy the same functional niche, which remains the same even in phylogenetically distant species. Therefore the identification of orthologues is more reliable for functional inference than comparing two paralogous sequences, which are similar without necessarily fulfilling the same biological role. Besides estimating sequence similarity for the purpose of functional transfer, the actual phylogenetic relationship between sequences is important as well. In addition to detecting sequence similarity and phylogenetic relationships, the possibility of convergence has to be accounted for in the homology approach.

Homologues of a given protein are normally found in a protein or DNA database by specialized tools, for example BLAST (Altschul, 1997) or FASTA (Pearson, 1988). As a simplistic approach, the function of the best scoring hit returned by such a search is transferred to the query sequence. This method can be refined by examining a larger number of hits that exhibit a certain degree of sequence similarity. Consideration of many hits in turn may include sequences, which only show regional similarity to the query. When trying to use those 'partial

homologies' for functional inference, it is a prerequisite that the information meant to be transferred from the hit to the query really resides within the matching region. By deducing a consensus from the classification of multiple reliable and preferably global hits, the query sequence can be assigned a specific function or a more general classification such as the involvement in a certain biological process. In this respect the usage of a unified functional vocabulary greatly facilitates the determination of a consensus classification (Ashburner, 2000).

1.1.2 The modular architecture of proteins

In the best case of a sequence-to-sequence comparison involving both the query and a clearly similar sequence, the region of similarity spans the complete length of the query. More frequently, the direct comparison reveals only a partial match between the two sequences. Strictly speaking, this situation allows only a functional classification of the particular stretch of the query sequence that was responsible for the reported database hit. One possible explanation might be a higher divergence of the sequences in the dissimilar region, but many of these constellations are caused by the modular architecture of proteins involved.

From the analysis of 3D protein structures it is known that a large portion of proteins contain multiple folding units rather than one monolithic fold. A folding unit, generally termed a domain, is a compact structure that folds independently from other parts of the protein. Typical domains have a hydrophobic core and consist of secondary structure elements such as β-strands or α-helices, which in turn can arrange themselves into sheets or a-hairpins. The exterior of a domain is usually hydrophilic due to its exposure to the solvent, but may also exhibit hydrophobic patches in order to fulfil certain functions, e.g. acting as a binding site. Within some multi-domain proteins, such binding sites are used to minimize the intramolecular repulsion of the domains. Others have their domains connected via flexible linkers. Another interpretation of a domain is that of an intra-protein subunit in analogy to the formation of a quaternary structure by separate proteins. As structure defines function, domains are generally associated with particular functions like enzymatic activity or ligand recognition.

In the course of evolution, the autonomous folding capabilities of domains seem to have made them suitable evolutionary units that could be arranged in different domain contexts without disturbing their structural integrity and therefore their function. An example for a domain, which exists as a monolithic protein as well as part of multi-domain proteins with different structures, is the ubiquitin-like domain (Figure 1-1). The underlying evolutionary events that lead to the formation of new domain organisations are mainly exon shuffling, duplication and fusion of whole genes or just gene regions (Li, 1997). Duplication of gene

regions may lead to a repetitive domain structure, whereas fusion or insertion can generate 'mosaic proteins' composed of domains originating from genes with different evolutionary histories.

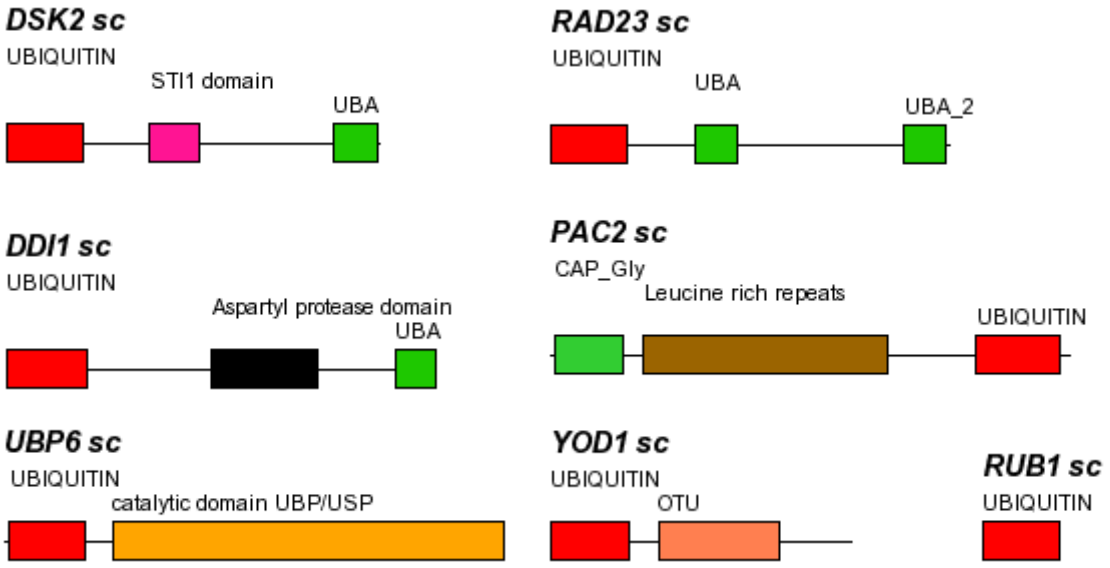


Figure 1-1 Domain topologies of multi-domain proteins with a Ubiquitin-like domain. In proteins like Ubiquitin, SUMO and Rub1/NEDD8 (shown here), this type of fold is able to form monolithic proteins. 'sc' following the protein name indicates *S. cerevisiae*.

1.1.3 Homology domains and their impact on protein classification

Common domains of otherwise unrelated protein sequences may be used for functional classification. It has to be kept in mind that this type of classification is restricted to the domains under consideration. Common domains from different proteins often share congruent boundaries and a homology relationship and are thus called 'homology domains'. In many cases these homology domains correspond to structural domains, which are thought to exhibit folding independence. Homologous regions shorter than approximately 20 residues are too small to form an independent hydrophobic core and thus should not be considered true 'domains'. Nevertheless, those small conserved regions can be carrier of important functional information, e.g. by being a recognition target for other proteins. In the following, short conserved regions are referred to as 'motifs' instead of 'domains'.

As many proteins are multi-domain proteins, the protein classification based on homology domains inevitably leads to more than one function for such proteins, because each homology domain can have its own characteristic function. Therefore these proteins belong to more than one protein family. As a consequence, the most accurate approaches to protein classification rely on domain-to-domain rather than on complete protein-to-protein comparisons. Another reason to compare domains individually comes from the observation that within mosaic proteins with the same set of domains, the domain organization may be shuffled.

Once a homology domain is identified as a sequence stretch conserved across various proteins and attributed to a certain function, it can serve as a template for the classification of novel sequence data. To that end, several techniques have been developed that aim at the extraction of the essential features of a homology domain, and store them in terms of motif descriptors ('profiles') (Bucher, 1996). This concept will be introduced in the following, as sequence profiles can be applied as a very sensitive method to find distant homologous members of a protein family and therefore are a central technique of this work.

1.2 The ubiquitin-proteasome system

Most of the homology domains used by proteins of the ubiquitin-proteasome system (UPS) described below are present exclusively in this pathway. Thus, novel proteins containing one of these UPS-specific domains may be considered as new components of the UPS with a high reliability. Indeed, in the recent past the mining of sequence databases for proteins with domains relevant to the UPS has been a valuable source for new components and regulators of this system (Bai, 1996, Hofmann, 1996, Hofmann, 1998, Hofmann, 2001).

Common to all eukaryotic cells is their capability to degrade proteins and peptides, and for this purpose two major proteolytic system, the 26S proteasome and the lysosome are present within the cells. While the lysosome is responsible for the non-specific degradation of endocytosed proteins such as receptors, the proteasome bears the main load of intracellular proteins to be degraded. The proteasome is a multi-subunit protease that combines substrate recognition, unfolding and hydrolytic cleavage (see Figure 1-2 for a rough overview of the UPS). Prior to proteasomal digestion, substrates are usually tagged with a poly-ubiquitin chain via a covalent isopeptide bond that links the free C-terminus of a ubiquitin (Ub) and the ε-amino group of a lysine. This multi-step enzymatic reaction is generally known as ubiquitylation.

Until Ub is linked to a substrate protein in a covalent manner, it passes through several enzymatic reactions. First, Ub precursors have to be processed to allow activation by Ub-activating enzymes (E1) (see chapter 1.2.1 and 1.2.2).

Secondly, the activated Ub is transferred from the E1 to a so-called Ub-conjugating enzyme (E2) (see chapter 1.2.3) (Hershko, 1983).

Then a Ub-ligase (E3) catalyses the transfer of the Ub moiety to a substrate via one of two major types of transfer mechanisms (Huang, 2004). E3 enzymes form a heterogeneous group of proteins belonging to different protein families and will be described in more detail in chapter 1.2.4. Enzymes that catalyse the elongation of Ub chains by ligating Ub to existing poly-Ub chains are often referred to as E4 enzymes (Koegl, 1999). They can be considered as

specialized E3s, and correspondingly share the same sequence motifs as the other E3s (Pickart, 2004).

Ubiquitylation is a reversible process, and eukaryotic genomes harbour a set of deubiquitylating enzymes of various evolutionary background. These enzymes use different homology domains to cleave both isopeptide bonds between Ub moieties in poly-Ub chains and protein-Ub conjugates and are described in chapter 1.2.5.

The proper attachment of Ub to a substrate requires a lysine-based ubiquitylation site and specific surface patches that are recognized by the substrate-binding site of an E3. The sequence features involved are diverse and hardly amenable to sequence analysis (Peters, 2002). In contrast, Ub recognition motifs are readily recognizable in multiple Ub binding proteins and seem to be widely applied throughout the UPS (see chapter 1.2.6) (Hofmann et al., 1996, Hofmann et al., 2001).

The multi-subunit proteasome itself is also characterized by recurring homology domains. For example, the cylindrical, proteolytic core particle of the proteasome (20S) consists of 28 homologous subunits. Moreover, the two subcomplexes of the 19S regulatory particle, base and lid, contain particular homology domains (Ferrell, 2000). In this respect, the base complex harbours six AAA-ATPases and the 'lid' consists of eight subunits stemming from two different protein families (Hofmann et al., 1998, Maytal-Kivity, 2002). Each protein family contributing to the 26S proteasome structure will be described in chapter 1.2.7.

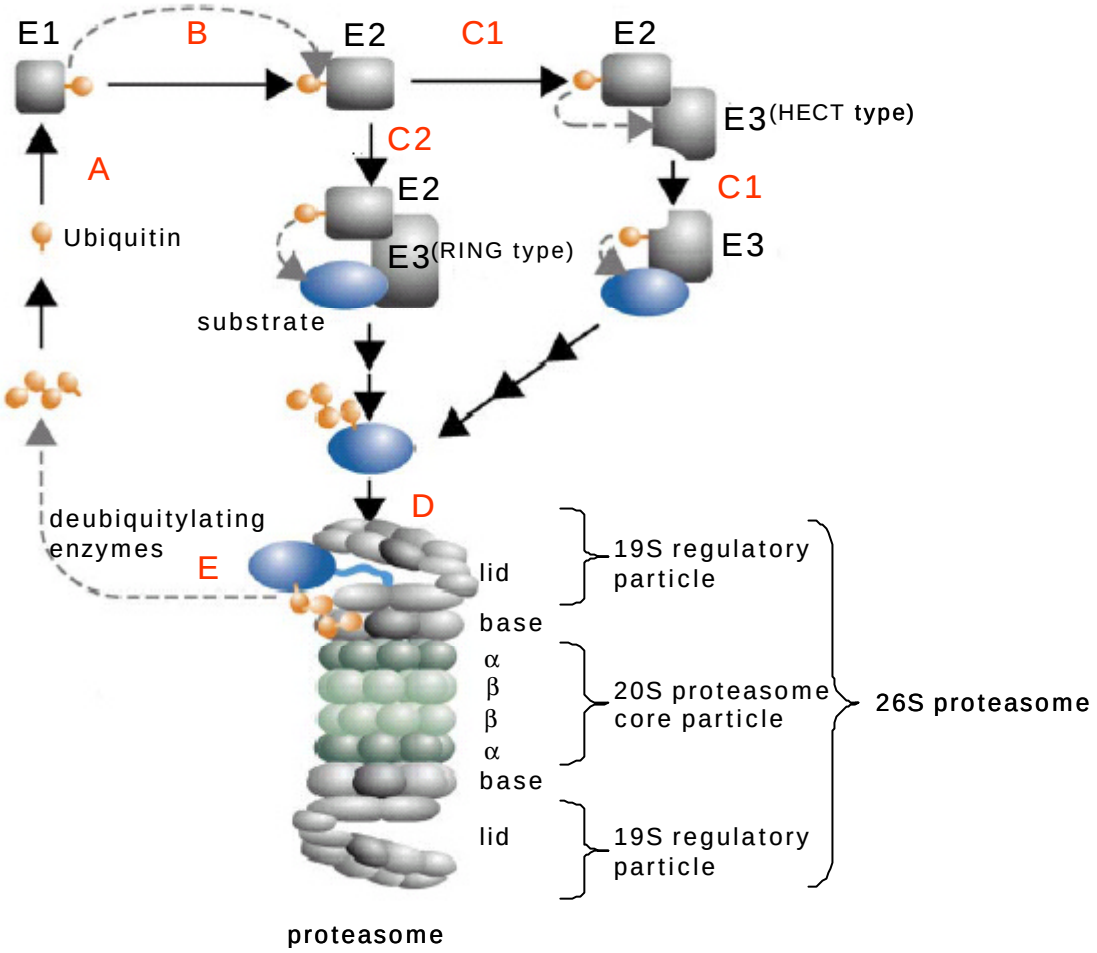


Figure 1-2 Simplified overview of the ubiquitin-proteasome pathway. A, activation of Ub by an E1; B, transfer of activated Ub to an E2; C1/C2, recognition of a substrate molecule by an E3 and biosynthesis of a substrate-linked poly-Ub chain; D, binding of the ubiquitylated substrate to the proteasome and substrate degradation; E, recycling of Ub by deubiquitylating enzymes for subsequent rounds of substrate ubiquitylation. More details will be given in the text. Figure adapted from Kloetzel, 2004.

1.2.1 Ubiquitin and its relatives

Ubiquitin

Ubiquitin (Ub) is a small protein of 76 residues and is ubiquitously found in all eukaryotic species. Its primary sequence is extremely well conserved and can easily be detected in quite different organisms. Ub is usually translated as a precursor, which consists of multiple in-frame fused Ub copies or of Ub fused to other highly expressed proteins like ribosomal subunits (Redman, 1994). Prior to their use in the UPS, Ub precursors have to be processed by Ub specific hydrolases (Amerik, 2000, Finley, 1989).

The primary role of ubiquitin as a degradation signal for proteins is achieved by its attachment to the substrate via a covalent isopeptide bond. The substrate may be Ub as well, leading to poly-Ub chains. So far, three lysine residues of Ub have been demonstrated as possible ubiquitylation sites (K29, K48 and K63) and different types of linkage seem to be associated

with different functions within the cell (Pickart, 2000). Even monoubiquitylation serves a specific role and was reported to be utilized as a signal for internalisation of receptors (Di Fiore, 2003, Terrell, 1998).

Type I Ub-like proteins

Ubiquitin has a multiplicity of homologous proteins that also can be attached to other proteins posttranslationally, e.g. SUMO family proteins (Melchior, 2003, Muller, 2001), NEDD8/Rub1 (Ohh, 2002), Urm1 (Goehring, 2003), FUBI/FAU (Michiels, 1993), Hub1 (Dittmar, 2002), ISG15 (Kim, 2003) and Fat10 (Raasi, 2001). The latter two consist of two fused Ub-like domains and have only been found in vertebrates so far. Besides the obvious Ub homologues, there are several analogous protein modifiers, e.g. Atg8 (Mizushima, 2003), Atg12 (Wang, 2003) and Ufm1 (Komatsu, 2004). Their relationship to the Ub family is not yet fully understood. Like Ub, at least some of these modifiers use cascades of activating and conjugating enzymes, as well as proteins recognizing and removing the modification of a substrate. Ub and proteins that can act as modifiers are generally referred to as type I Ub-like proteins or Ub-like modifiers.

Except for Hub1, Ub and its homologous type I Ub-like proteins end with a "GG" motif. As the "GG" motif has been discussed as a prerequisite for conjugation to substrates (Jentsch, 2000, Rudolph, 2001), Hub1 probably requires mechanisms different from that of the "GG" motif containing type I Ub-like proteins. Whether Hub1 is covalently attached to other proteins is still controversial (Luders, 2003). Despite the described similarities, type I Ub-like proteins typically do not mark their substrates for proteasomal degradation. For example, SUMO-conjugation targets cytosolic RanGAP1 to the nuclear pore complex (Matunis, 1996) and SUMOylation of p53 leads to its activation (Gostissa, 1999). NEDD8/Rub1 is conjugated to the Cullin subunits of SCF complexes in order to regulate their activity (Lammer, 1998, Ohh et al., 2002) and was found to modify p53 (Xirodimas, 2004). Fat10 was reported to be conjugated to so far unknown proteins and to stimulate apoptosis (Raasi et al., 2001). Unlike Ub but similar to Nedd8, Fat10 is a substrate of the proteasome (Hipp, 2004).

Type II Ub-like proteins

Type II Ub-like proteins contain a Ub-like homology domain, but are not conjugated to substrates (Jentsch et al., 2000). The Ub-like domain lacks the C-terminal "GG" motif, which is a hallmark of type I Ub-like proteins and a likely prerequisite for attachment to other proteins. Proteins with a Ub-like domain often are mosaic proteins containing other, UPS associated homology domains; some examples are shown in Figure 1-1. Prominent examples for type II Ub-like proteins are yeast Rad23 and its human orthologues Rad23A and Rad23B, which in

addition to the Ub-like domain contain two Ub-binding UBA domains. These proteins play an important role in nucleotide excision repair of DNA, and the Ub-like domain is necessary for this function (Prakash, 2000, Watkins, 1993). Furthermore, the Ub-like domain of Rad23B was shown to associate with the S5a subunit of the human 26S proteasome (Hiyama, 1999).

Another widespread Ub-like domain is the UBX domain, which shares the same fold with Ub and plays a role in the UPS as well (Buchberger, 2001). For example, fission yeast Ubx2 and Ubx3 both contain a UBA and a UBX domain with the UBX domain mediating interaction with the hexameric p47/VCP/Cdc48 complex (Hartmann-Petersen, 2004).

1.2.2 Ub-activating enzymes (E1)

Before free Ub enters the ubiquitylation machinery, it is activated by enzymes termed E1 (Ub activating enzymes). The activation is ATP-dependent and is subdivided into two steps, both catalysed by an E1. First, the C-terminus of the free Ub or Ub-like protein becomes adenylated. In the second step, the activated C-terminus is transferred to the catalytic cysteine residue of the E1 yielding a highly energetic thioester bond (Pickart et al., 2004). Besides activation, E1s bind to Ub conjugating enzymes (E2), which are downstream components of the ubiquitylation process, and transfer the Ub moiety to the catalytic cysteine of these enzymes afterwards.

All E1s known so far share a common homology domain, which contains a NAD binding site. This domain assumes a fold found in many NAD binding proteins and can be traced back even to bacteria. Here, it is detected in proteins of thiamine and molybdopterin biosynthesis pathways, ThiF and MoeB, respectively (Begley, 1999, Unkles, 1999). These proteins catalyse the adenylation of the C-termini of ThiS and MoaD, two proteins with structural similarity to Ub (Lake, 2001, Rudolph et al., 2001), but do not transfer them to proteins. Rather, activation of ThiS and MoaD serves for sulphur transfer in the corresponding biosynthetic pathways (Pitterle, 1993, Taylor, 1998).

Interestingly, E1s harbouring two NAD binding domains like yeast Uba1, human UBE1 or human UBE1L are active as monomers, while those with a single copy fulfil their function only in complex with a protein that have a second NAD-binding domain (Huang et al., 2004). This may be a copy of the same protein as it is the case for Atg7, or a distinct homologue, as found in the pairings of yeast Aos1/Uba2, human Aos1/Uba2 and human APPBP1/Uba3 (Johnson, 1997, Komatsu, 2001, Walden, 2003).

E1s exhibit a modular architecture as seen in several solved E1 structures (Lois, 2005, Walden et al., 2003). Besides the common NAD binding domain, in some E1s a Rhodanese

domain is present, e.g. in Uba4/MOCS3. This observation is of particular interest, as the Rhodanese is also found in bacterial ThiI, a protein involved in the thiamine biosynthesis pathway (Palenchar, 2000). The Rhodanese domain is known to have sulphur transferase activity, but its role in some E1s is so far unknown. Another example for the modular architecture of E1s is the Ub-like domain recently reported in the human SUMO-E1 Sae2 (Lois et al., 2005). A detailed analysis of E1 domain structures will be given in chapter 3.

1.2.3 Ub-conjugating enzymes (E2)

After a thioester bond is established between Ub and an E1, the latter recruits a second class of enzymes important for ubiquitylation, dubbed E2s or Ub-conjugating enzymes (UBC). Once recruited, the E2 itself becomes ubiquitylated itself in a transthiolation reaction (Pickart, 1985). Interestingly, there are multiple Ub-specific E2s known (11 in yeast) (Pickart et al., 2004), while for the enzymatic cascades of the modifiers SUMO/Smt3 or NEDD8/Rub1 only one E2 has been discovered so far, which is Ubc9 or Ubc12, respectively (Hershko et al., 1983, Johnson, 1997, Pickart et al., 1985, Schwarz, 1998). The function of an E2 is not necessarily restricted to one particular modifier, as seen in the case of human UBE2E2/UCH8. The latter was recently reported to conjugate Ub as well as the linear di-Ub-like ISG15, indicating at least in this case overlapping pathways of Ub and a Ub-like modifier (Zhao, 2004).

Independent of the modifier conjugated, all E2s have a conserved homology domain termed UBC (Ub conjugating) in common, which is ~150 residues in length and harbours the catalytic cysteine (VanDemark, 2002). Besides this domain, some E2s have large sequence extensions up- and downstream of this domain, which in some cases play a role in E3 recognition (Mathias, 1998). A striking example for an E2 bearing much primary sequence outside the common homology domain is the ~5000 aa BRUCE/BIRC6 (Hauser, 1998).

E2s act on Ub function in distinct biological processes. One very specific and important function is performed by the yeast E2 Cdc34, which is responsible for the degradation of cyclin-G1 and Sic1, two key regulators of the cell cycle (Blondel, 1996). Yeast Rad6 has been shown to be essential for degradation of N-end-rule pathway substrates as well as for modification of histones and the polymerase processing factor PCNA (Dohmen, 1991, Dover, 2002, Hoege, 2002). Human UBEL1 and UBE2E2 are interferon inducible E2s important for ISG15 conjugation and therefore play a role in immune response (Kim, 2004, Zhao et al., 2004). More E2s and their specific function are reviewed by Haas and Siepmann (Haas, 1997).

In rare cases, the E2 can directly transfer the modifier to some substrates, but most often a specificity factor for substrate recognition is required. Another mode of modifier transfer

involves an intermediate step, in which the modifier is transferred to the specificity factor before final conjugation to the substrate. The specificity factor is in both cases termed E3 or Ub-ligase.

1.2.4 Ub-ligases (E3)

Ub-ligases (E3) have the function of recognizing a substrate and mediating the transfer of Ub to the substrate. Organisms generally possess a large number of E3s, each responsible for a limited set of substrates. Thus, E3s provide specificity in substrate ubiquitylation. In the case of Ub, E3s are a prerequisite for substrate recognition, while SUMO can be transferred to some substrates in the absence of an appropriate E3 (Hershko, 1998, Seeler, 2003). There is evidence that the E2 for SUMO/Smt3, Ubc9, directly interacts with the substrate RanGAP1 via binding to a sumoylation consensus site hKx(D|E) (Bernier-Villamor, 2002). In this regard, Ubc9 acts as a combined E2/E3 enzyme taking over the activated SUMO from its E1 and transferring it to the substrate. Nonetheless, several SUMO-specific E3s have been reported, which act as bridging factors like Ub-E3s bringing both E2 and substrate into a sterically favourable arrangement (Dohmen, 2004).

So far, no universal recognition motif for ubiquitylation comparable to the sumoylation consensus site is known and this observation is likely associated with the large number of Ub-specific E3s in the UPS. One exception might be the 'N-end-rule-pathway' that is responsible for the proteasome dependent degradation of proteins with a destabilizing N-terminal residue (F, H, I, K, L, R, T, W), which may be regarded as a degradation signal ('degron'). Based on sequence analysis, the E3 components for the ligation of Ub can be subdivided into three major classes (HECT, RING and U-Box), but a fourth protein family characterized by a particular homology domain (A20 zinc finger) has recently joined the ranks of ubiquitin ligases (Deshaies, 1999, Jiang, 2001, Scheffner, 1990, Wertz, 2004).

HECT based Ub-ligases

Proteins of the HECT family share a C-terminal homology domain of approximately 350 residues. The first HECT family member discovered to have ligase activity was E6AP. Therefore, the homology domain was named after this protein, 'homologous to E6AP carboxyl terminus' (Scheffner, 1995). E6AP has been shown to bind and ubiquitylate the tumour suppressor p53 in cells infected with the human papilloma virus, leading to proteasomal degradation of p53. One of the natural targets of E6AP is human Rad23A (Kumar, 1999).

The HECT domain binds the E2 enzyme and contains the catalytic cysteine that gets linked to Ub in a transthiolation reaction between a ubiquitylated E2 and a HECT E3 (Pickart, 2001, Scheffner et al., 1995). HECT type E3s are thought to transfer a poly-Ub chain onto the

substrate at once, with the poly-Ub chain being assembled first on the HECT E3 (Pickart et al., 2004).

RING-based Ub-ligases

The second class of E3s is the so-called RING-finger family (for 'really interesting new gene'). RING finger proteins share a globular domain, whose structure depends on complexing two zinc ions. Unlike the HECT proteins, RING finger proteins do not get covalently linked to Ub, but rather function as adaptors between the substrate and the E2. Another characteristic feature of the RING family is that some members are subunits of large multi-subunit E3 complexes like the SCF-complex or the APC (Deshaies, 1999, Peters, 2002). Other RING finger proteins, like p53-ubiquitinating Mdm2, work without auxiliary proteins (Li, 2003).

Some Ub ligases use a U-box for E2 recruitment. The U-box represents a highly divergent variant of the original RING finger motif (Aravind, 2000, Pringa, 2001). A well known U-box protein is CHIP, which associates with chaperones like Hsc70 or Hsp90 in order to recognize the substrates to ubiquitylate, e.g. the glucocorticoid receptor (Connell, 2001, Cyr, 2002).

The SCF and other RING-cullin-based Ub-ligase complexes

The SCF complex (for 'Skp1, Cullin, F-box') is a multisubunit Ub-ligase whose core consists of the RING finger protein Hrt1/Roc1/Rbx1, the cullin Cdc53 and Skp1. Analogous to the monomeric RING finger E3s, the RING finger domain of Hrt1 is utilized for E2 recruitment. Cdc53 serves as a scaffold binding the Hrt1 subunit and Skp1 simultaneously (see Figure 1-3 B). Skp1 in turn ties the real substrate binding protein, which contains two major domains, an F-box domain utilized for Skp1 binding and a further protein interaction domain. For example, the yeast F-box protein Cdc4 contains a repetitive WD40 region that adopts a β-propeller fold suitable for binding the substrate, Sic1 (Deshaies, 1999). The F-box subunit may be regarded as an exchangeable substrate specificity factor, thereby allowing the SCF to ubiquitylate different targets while the E3 core remains unchanged. From the structure of the SCF a more general model was developed valid for similar types of complex E3 ligases (see Figure 1-3 A). The main features that differ between the SCF complex and related complexes is the usage of completely different substrate specificity adaptors like SOCS-box or BTB proteins (see Figure 1-3 C,D) (Willems, 2004). The APC ('anaphase promoting complex') also belongs to the family of complex SCF-type E3 ligases, but uses a distinct RING finger protein, Apc11, and contains markedly more subunits (Zachariae, 1998).

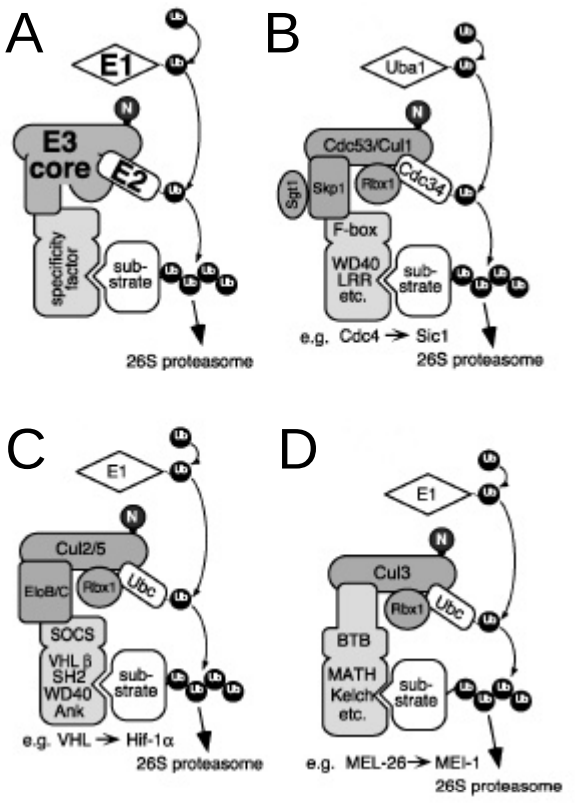


Figure 1-3 Cullin-RING-based E3 complexes (SCF-like complexes); A, general composition of a cullin-RING-based E3 complex; B-D, specific examples together with an example adaptor and its corresponding substrate are shown. See 1.2.4 for more details. Figure adapted from Willems (Willems et al., 2004).

1.2.5 Ub-hydrolases (DUB)

Deubiquitylating enzymes (DUBs) form a heterogeneous enzyme group, whose members cleave ubiquitin-linked proteins after Gly76, the terminal ubiquitin residue (Lam, 1997). These enzymes can participate in two different cellular processes, biosynthesis of free Ub and deubiquitylation of Ub-protein conjugates.

A family of DUB enzymes that preferably cleaves Ub monomers from Ub precursors consists of small thiol-proteases (~ 25 kDa) and is often referred to as UCH family (Ub C-terminal hydrolases). They cleave regular peptide bonds at the C-terminus of Ub in poly-Ub precursors or in fusion proteins consisting of ribosomal proteins and Ub. However, there are examples of UCH-type DUBs that act both on precursors and conjugates (Kwon, 2004).

Unlike the processing of precursors, the deubiquitylation of Ub-protein conjugates requires hydrolysis of an isopeptide bond. The family of Ub specific proteases (USP in human, UBP in yeast), which is larger than the UCH family, and whose members are larger in size (60-300 kDa), prefers the cleavage of such isopeptide bonds. Only for a few USPs functions are known, e.g. USP11 has been shown to deubiquitylate RanBPM (Ideguchi, 2002) and UBP3 has been implicated in gene silencing (Moazed, 1996). There seems to be a broad diversity

concerning their function, their dependence on ATP and their localization within the cell. Some USPs occur freely, while others are associated with large complexes, such as the proteasomal lid, the CSN or the SAGA complex (Daniel, 2004, Leggett, 2002, Zhou, 2003). The proteasome-associated USPs are thought to be responsible for protecting Ub from proteasomal degradation. Other USPs like yeast Ubp14 and human UCHL5/UCH37 probably play a role in editing poly-Ub chains on target proteins and for Ub recycling by cleaving free poly-Ub chains instead (Amerik, 1997, Lam, 1997, Lam et al., 1997).

Ataxin-3, the protein mutated in Machado Joseph Disease (SCA3), belongs to a novel group of cysteine-proteases and is active against ubiquitin chains (Burnett, 2003, Scheel, 2003). Like ataxin-3, the OTU (ovarian tumour) proteases display a structural similarity to the USP protein family in their catalytic core (Makarova, 2000) and a deubiquitylating activity was shown for several OTU proteins (Evans, 2003, Soares, 2004).

Besides these four DUB classes of cysteine proteases, a deubiquitylating activity was found in the MPN subunit Rpn11 of the proteasomal lid (Maytal-Kivity et al., 2002, Verma, 2002, Yao, 2002). Rpn11 generates free poly-Ub chains by hydrolysing the bond that connects the target protein and the proximal Ub of the poly-Ub chain. Interestingly, a deneddylating activity, i.e. the cleavage of Lys-linked Nedd8 conjugates is intrinsic to the MPN protein Csn5, the CSN subunit analogous to Rpn11 (Cope, 2002, Maytal-Kivity et al., 2002). In contrast to UCH and UBP proteases, MPN proteins are metalloproteases coordinating Zn^{2+} in their active site (Tran, 2003, Verma et al., 2002, Yao et al., 2002).

1.2.6 Ub-binding proteins

Typical intracellular signal transduction pathways are characterized by a modular architecture of the proteins involved in the three fundamental steps of signal generation, signal recognition and signal removal. In protein phosphorylation, the archetype of such transduction systems, the three roles are filled by kinases, phosphatases, and phosphopeptide recognition domains (SH2, PTB, FHA etc), respectively. The components of the UPS obviously form an analogous system. Here, the E1-E2-E3 cascade corresponds to the signal generation, where Ub conjugated to a substrate constitutes the signal itself. Ub-binding proteins serve for recognizing the signal, while DUBs quench it. Similarly, the analogy to signal transduction pathways seems to be valid for most Ub-like modifiers and their associated apparatus.

Within the UPS, there exist many recognition systems, which can recognize the different ubiquitylation states including different types of Ub-to-Ub linkages and various chain lengths. Ub-recognizing proteins are normally classified according to the homology domains involved and play a crucial role in the UPS.

UBA domain

The Ub associated domain (UBA) occurs in many different proteins of the UPS, including E3s, DUBs, Ub conjugases and adaptors (Hofmann et al., 1996). The universal character of the UBA as a Ub binding domain can be seen from many reported interactions between UBA containing proteins and Ub (Bertolaet, 2001, Rao, 2002, Wilkinson, 2001).

The UBA domain is a small domain of only ~40 residues with a three-helix bundle fold (Mueller, 2002). It has a preference for tetra-Ub chains, which is of two orders of magnitude higher than to mono-Ub (Wilkinson et al., 2001). There are contradicting reports on the linkage preference of UBA domains. Both a binding to Lys-48 linked chains and to Lys-29 linked chains have been described (Raasi, 2003, Rao et al., 2002).

While the uncertainty on linkage preference of UBA domains remains, more information exists on the part of Ub that is recognized by UBA domains. By NMR-based methods, Ryu et al. have identified the Ile-44 surface patch of Ub, and a homologous region in the Ub-like domain of human Rad23B, as interacting with UBA domains (Ryu, 2003). These experiments also demonstrate the ability of UBA domains to interact with type II ubiquitin-like proteins.

CUE domain

The CUE domain is another ubiquitin-binding homology domain (Ponting, 2000), which has been suggested to be distantly related to the UBA domain (Shih, 2003). This relationship was recently confirmed by Kang et al., who have solved the Cue2 structure in complex with Ub (Kang, 2003). In this CUE/Ub complex, the CUE domain is bound to Ub's Ile-44 patch, similar to the binding of the human Rad23B-UBA domain to Ub (Ryu et al., 2003). Additional evidence for the CUE domain as a Ub binding domain comes from Donaldson et al. and Shih et al., who have reported the CUE domain of yeast Vps9, Cue2, Cue3, Cue5 and human Tollip to directly bind mono-Ub (Donaldson, 2003, Shih et al., 2003). The preference for mono-Ub is probably valid for all CUE domain proteins and makes it different from the UBA domain, which prefers poly-Ub (Shih et al., 2003). Another CUE family member, Cue1, has been assigned a role in the ER associated degradation pathway (ERAD), which relies on Ub signals (Biederer, 1997). However, its affinity to ubiquitin is significantly reduced compared to Vps9 and Cue2 (Shih et al., 2003).

UIM

The UIM (Ub interacting motif) was first described in 2001 by Hofmann et al. based on a motif in Rpn10/S5a (Hofmann et al., 2001), a proteasomal subunit, which had been shown to bind ubiquitin (Young, 1998). Other proteins associated with the UPS contain this motif as well, e.g. several DUBs, the yeast F-Box protein Ufo1 and some E3s. Besides the UPS, the UIM

appears in proteins that regulate Ub-dependent events of endocytosis. Mono-ubiquitylation of target proteins serves as an internalisation signal, and the Ub-recognizing element in this process has been narrowed down to the UIM in eps15 and Hrs (Di Fiore et al., 2003, Polo, 2002). Interestingly, mono-ubiquitylation of endocytosis components such as ligand-bound receptors in the plasma membrane depends on a functional UIM domain in the same protein (Klapisz, 2002, Polo et al., 2002) and the UIM also keeps the ubiquitylation status of a target on mono-ubiquitylation (Di Fiore et al., 2003).

The UIM is a very short motif of ~20 residues consisting of an α-helix with the conserved residues located on one side of the helix (Shekhtman, 2002). Like the UBA domain, the UIM binds to the Ile-44 patch of Ub. This interaction of a UIM and Ub does not involve the Lys-48 of Ub, which would allow the UIM to differentiate between poly-Ub and mono-Ub. Nevertheless, within the UPS, the UIM obviously prefers poly-Ub as a binding partner (Perez, 2003, Polo et al., 2002, Shekhtman et al., 2002, Thrower, 2000).

GAT domain

The GAT domain (GGA and Tom1) was initially found in proteins regulating clathrin-mediated trafficking of vesicles (Dell'Angelica, 2000). Recent findings have shown the GAT domain to bind to Ub (Katoh, 2004, Shiba, 2004). The structures of several GAT domains have been solved, presenting the GAT domain as a three-helix bundle with elongated and almost parallel helices, an arrangement quite different from the helix bundle of the UBA structure (Shiba et al., 2004). Like the UBA domain, the GAT domain is thought to interact with the Ile-44 patch of Ub (Shiba et al., 2004).

UEV domain

The UEV (Ub E2 variant) domain is related to the domain responsible for the catalytic E2 activity, but is devoid of the cysteine important for Ub conjugation (Ponting, 1997). A well known member of this inactive subfamily of E2 enzymes is the tumour susceptibility gene 101 protein (TSG101/Vps23), which plays a role in Ub-dependent protein sorting and is mutated in certain types of breast cancer (Bishop, 2002, Pornillos, 2002, Pornillos, 2002). Budding yeast Mms2, another UEV protein, forms a complex with Ubc13 (functional E2) and is required for Rad6/Rad18 dependent postreplicative DNA repair (Broomfield, 1998, Hofmann, 1999). Available structural information on TSG101 in complex with ubiquitin demonstrates the ability of the UEV domain to bind Ub (Sundquist, 2004). At the same time, other UEV proteins may differ in their Ub binding modes (Sundquist et al., 2004).

NZF domain

The NZF domain (*N*pl4 *Z*n-*f*inger) is a C4-type Zn-finger that coordinates a zinc ion via four cysteines, which makes this domain very different from the Ub-binding modules described so far (Wang, 2003). This Zn-finger is found in human Npl4, a VCP/Cdc48/p97 adaptor protein, and in yeast Vps36. Experiments with both proteins revealed Ub-binding properties (Alam, 2004, Meyer, 2002).

1.2.7 Proteasome

20S proteasome

The 20S proteasome is a subcomplex of the 26S proteasome and after binding of two copies of the 19S regulatory particle yields the 26S proteasome. The barrel-shaped 20S proteasome consists of 28 subunits arranged in four stacked rings with seven subunits each. All subunits share a common evolutionary ancestor and can be further subdivided into the α-subunits forming the outer rings and the β-subunits, which are found in the two inner rings. This α7β7β7α7 structure of four rings harbours three major chambers in its centre. All chambers are connected with each other and the surrounding solvent. The largest and centrally located chamber bears the six catalytically active sites, which are located on distinct subunits of the beta-rings. The active subunits are termed β1, β2, and β5, each of which occurs with two copies in the 20S proteasome. The sequestration of the protease activity to the shielded chamber allows the proteasome to limit degradation to the correct substrates. Before a substrate can be degraded within the central chamber, it has to be recognized as a correct substrate at one of the entry pores and unfolded in a subsequent step (Baumeister, 1998).

Archeae and several bacteria also possess proteasome-like proteases, which typically only consist of one or two subunit types. During evolution, the number of distinct subunits has multiplied in higher organisms, i.e. the yeast genome encodes seven different α-subunits and seven β-subunits (Gille, 2003). The situation in mammals is even more diverse with ten distinct β-subtypes, three of which are interferon-γ inducible ('immunosubunits') and only found in so-called 'immunoproteasomes' (Kloetzel, 2004). The three immunosubunits β1i, β2i and β5i occupy the positions of β1, β2 and β5 positions, respectively, of newly synthesized 20S immunoproteasomes upon interferon-γ induction. As a consequence of the altered catalytic subunit layout, immunoproteasomes generate peptides suitable as antigen precursors. The latter normally have to be trimmed to the correct length by cytosolic or ER-based peptidases (Kloetzel, 2004). Suitable antigens are finally presented by MHC class I proteins. Interferon-γ

also stimulates the biosynthesis of PA28α and PA28β proteins, which assemble into the heptameric PA28 regulatory complex able to cap the immunoproteasome.

All α-subtypes are non-catalytic in nature and fulfil a regulatory function instead. Their N-terminal extensions, especially that of the α3 subunit, lock the pores of the 20S proteasome. Only upon binding to the 19S regulatory particle (PA700) or the PA28 complex in the case of the immunoproteasome, the α-subunits' N-termini become delocalised and open up the pores (Groll, 2000, Kloetzel, 2004).

19S regulatory particle

The 19S regulatory particle is essential for proteasomal activity and consists of two subcomplexes. The one binding the 20S proteasome is the 'base', which is a hexameric ring of AAA-ATPases with chaperone activity (Braun, 1999) and three additional subunits, Rpn1, Rpn2 and Rpn10. Base subunit Rpt5 and Rpn10 have been shown to bind to poly-Ub chains and therefore may function as receptors of ubiquitylated substrates (Deveraux, 1994, Lam, 2002). A more indirect role in substrate delivery to the proteasome has been mapped to Rpn1, as this protein associates with the adaptor protein Rad23, which in turn is responsible for the recognition of many ubiquitylated substrates (Elsasser, 2002).

The other subcomplex called the 'lid' has a more complex structure based on eight core subunits (Glickman, 1998). The lid is composed of multiple subunits harbouring the PCI domain, named after the three similar complexes ((i) proteasome lid, (ii) COP9 signalosome or CSN complex, (iii) eukaryotic translation initiation factor eIF3) that contain this domain (Hofmann et al., 1998). These complexes in turn are termed PCI complexes. Other subunits of the lid are characterized by a second homology domain called MPN (Mpr1-Pad1 N-terminal) (Aravind, 1998, Kapelari, 2000). The lid complex contains an intrinsic deubiquitylating activity, which is encoded by the MPN subunit Rpn11 that has the hallmarks of a metalloprotease as described in chapter 1.6.5 (Maytal-Kivity et al., 2002, Verma et al., 2002, Yao et al., 2002). Interestingly, Csn5, an MPN-bearing subunit of another PCI complex, the signalosome, also encodes a metalloprotease that can cleave Ub from proteins (Groisman, 2003). In addition, Csn5 is needed for the cleavage of the Ub-like protein Nedd8 from cullins (Cope et al., 2002). No specific function has been described for the PCI subunits of the lid so far.

1.3 Detection of homologues and protein family analysis

1.3.1 Sequence comparison methods

As homology serves as a vector along which information can be transferred from one known sequence to a new sequence, methods have developed that help deciding if two sequences are sufficiently similar to infer a relationship. The basic concept is the alignment, in which two or more sequences are arranged along each other, so that evolutionary or structurally equivalent residues are opposed. Multiple sequences can be aligned simultaneously as well, leading to multiple sequence alignments ('MSA'). MSAs are not only helpful to illustrate evolutionary events within a protein family, but also constitute the basis for phylogenetic tree construction, secondary structure prediction, homology modelling and, of special interest, the identification of conserved domains. With regard to the latter, homology domains and motifs often appear as columns with obvious conservation, while the adjacent primary sequence has a higher level of variation. When looking for further proteins with a given domain, information on amino acid frequency at each position within the domain was shown to be useful for so called profile-based techniques (Bucher et al., 1996, Gribskov, 1987). A more detailed description on profile construction as well as database search methods will be given in chapter 2.

For calculating a mathematically optimal alignment of two or more sequences, dynamic programming algorithms are typically used, for example the Smith-Waterman algorithm that looks for the best alignment between two subsequences ('local alignment') (Smith, 1981). These algorithms aim at maximizing the alignment score under an additive scoring scheme by incorporating as many positively scoring residue pairs as possible into the alignment. At the same time, negatively scoring pairs from unconserved residues and special penalties for gaps have to be minimized. Finally, a single optimal alignment and its corresponding score are reported. Efforts to increase the sensitivity of this type of comparison aim at scoring alternative alignments as well and weighting them by a probability value (Bucher et al., 1996).

1.3.2 Multiple sequence alignments (MSA)

MSAs are simultaneous alignments of more than two sequences (see Figure 1-4). As mentioned above, MSAs provide information on amino acid composition at each individual position within the alignment. Unfortunately, the calculation of a mathematically optimal MSA is computationally too expensive to be feasible (Wang, 1994). Current MSA generating programs rely on heuristics and none of them guarantees to report the fully optimised MSA. The programs used in this work, ClustalW and T-Coffee (Chenna, 2003, Notredame, 2000), belong

to the category of progressive MSA algorithm. In general, these algorithms assemble an MSA by adding sequences one by one. First, all pairwise distances are calculated, followed by a phylogenetic tree construction using the neighbour joining method. The phylogenetic hierarchy defines the order by which the sequences are added to the MSA. At the same time, a weighting of each sequence takes place, with the weight being proportional to the amount of unique evolutionary information the sequence contributes to the MSA (Altschul, 1989, Lipman, 1989). The MSA construction starts with the closest related sequences and then the less related sequences are added according to the tree. While the first sequence pair is aligned in a conventional pairwise alignment fashion, every further sequence is compared with a consensus of each position in the previous alignment. An improvement of this progressive algorithm is found in T-Coffee, which applies position-specific scoring schemes instead of a global substitution matrix and which can combine global and local alignments.

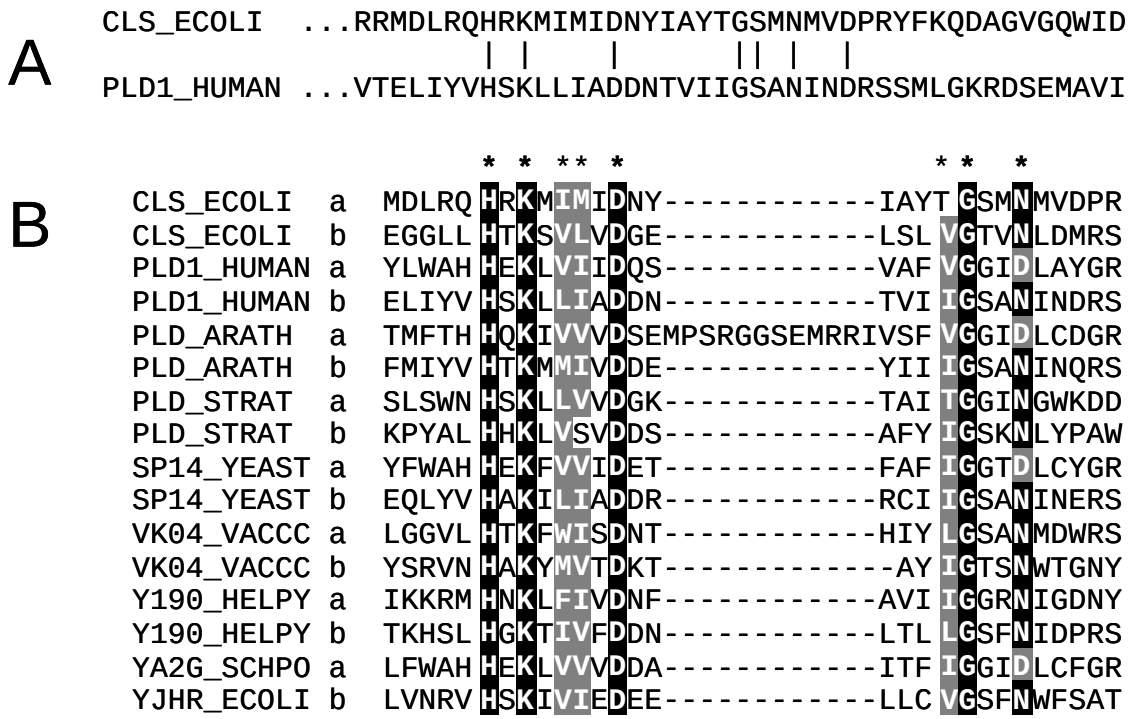


Figure 1-4 Multiple alignments reveal conserved positions. A, local alignment of two enzymes catalysing similar, phospholipase D (PLD) type reactions. The overall similarity within this alignment is ~10%; B, same sequences as in A now embedded in a multiple alignment together with representative members of the PLD family. Positions invariant or occupied with similar residues in at least 80% of the sequences are shown on black or grey background, respectively. 'a' and 'b' indicate the first and second copy of this motif in the sequences.

1.3.3 Profile searching

Standard sequence alignment methods have a common property: they use the same scoring matrix and gap penalties for all positions within an alignment. This behaviour can be traced back to the fact that during the comparison of two sequences no information on

conservation of individual positions or on the likelihood of gaps exists. In 1987, Gribskov introduced the profile technique as an extension to the Smith-Waterman method (Gribskov et al., 1987).

Profiles are position-specific scoring schemes

Profiles are derived from alignments of homologous proteins and represent a mathematical descriptor for all kinds of homology regions. The basic idea behind the profile technique is to treat the positions of an alignment as non-equivalent.

Each position of the profile stores an array of 20 score values, one for each amino acid that the profile might encounter when aligned to a new sequence. These position-specific scores are generated from the amino acids found in the corresponding alignment column, typically by applying a BLOSUM-type substitution matrix to each of the observed residues and summing up the results. As a consequence, a strongly conserved alignment column will yield a high positive score for the over-represented amino acid, and strongly negative scores for the under-represented or even absent residue classes. By contrast, a non-conserved alignment column will result in a relatively 'flat' profile position, with a weakly negative score for all of the possible amino acids. To summarize, position-specific score parameters reflect evolutionary conservation for each alignment position. The actual profile is a matrix, where the number of rows corresponds to the alignment length and the number columns corresponds to the 20 amino acids (Hofmann, 2000).

Gaps and other features can be included in the profile

Similar to the match scores, gap penalty parameters can be stored for each individual alignment position or transition between two alignment positions, respectively. Gap penalties can be adjusted to the occurrence of known insertions or deletions. If a particular insertion already has been observed in the initial alignment, it is obvious to assign lower gap penalties to that position. In the same manner, more expensive gap penalties are used in uninterrupted regions. Moreover, distinct positions in the alignment can be weighted differently, which in its extreme formulation can lead to the absolute requirement of a match at a certain position that, for example, harbours a catalytic residue. The type of profile used in this work, called ' generalised profile', also allows to anchor a domain or motif to the beginning or the end of a sequence.

Profiles can be used for database searches

A profile can be aligned to a sequence or a sequence database using dynamic programming algorithms. In contrast to sequence comparison methods described so far, no explicit substitution matrix is needed at that step, as the profile already contains such

evolutionary information in a processed form. The 'generalised profile' format allows different alignment modes, i.e., if the profile has to be matched entirely ('domain global') or partially ('local') to the target sequence (Bucher et al., 1996).

Iterative refinement

The sensitivity of profiles and their ability to discriminate between true matches and randomly occurring ones can be enhanced markedly by iterative refinement (Tatusov, 1994). In this process, database sequences found to be significantly related to the initial query profile can be used to augment the initial multiple alignment, which in turn can be used for the calculation of an improved 'second-round' profile. In general, the iterative refinement of profiles leads to increased sensitivity, but holds some problems as well. Major problems are the integration of similar, but non-homologous sequences and the treatment of large protein families.

Profile methods

The most popular methods used for profile searching are the 'generalised profile' technique of PROSITE, HMMER/Pfam and PSI-BLAST (Altschul et al., 1997, Eddy, 1998, Hulo, 2004). These databases and the construction of profiles will be described in chapter 2.

1.3.4 Substitution matrices

Conventional alignment methods calculate an optimal alignment between two sequences by application of a scoring scheme. These schemes consist of a 'substitution matrix', which contains scores for all possible residue pairs, as well as penalty scores for insertions and deletions. Existing scoring schemes mainly have been determined by probabilistic means from the analysis of sequences known to be related. Substitution matrices quantify preferences for certain amino acid substitutions over others during evolution. Besides abundance of each amino acid, the evolutionarily derived scoring schemes often reflect physicochemical similarities between the amino acids. Pairwise and multiple sequence alignment methods, as well as profile techniques, rely on those scoring schemes, which therefore play a crucial role in sequence analysis. In the course of this work a substitution matrix called BLOSUM45 was generally applied. BLOSUM45 is a particular matrix of the BLOSUM (BLOCKS substitution matrix) matrix series. It is directly computed from a MSA of true protein families by counting the substitution events. The '45' indicates a minimum of 45% identity between sequences of the source MSA (Henikoff, 1991).

1.3.5 Dendrogram analysis of proteins and genes

Definition of dendrograms

Originally, a phylogenetic analysis aims at resolving the evolutionary relationship among organisms based on a particular protein or nucleotide sequence. These relationships can be illustrated using tree-like diagrams ('phylogenetic trees', 'gene trees', 'dendrograms'), in which branches indicate 'evolutionary time' or simply evolutionary relationships and nodes as well as leaves represent a gene or protein sequence at a certain time point in evolution. In this respect, external nodes correspond to contemporary sequences and are called 'operational taxonomic units' (OTUs), while internal nodes either reflect gene duplication or speciation events. Trees can be either rooted if the direction of time is known or unrooted if not. Rooted trees are anchored to a special internal node, called the root, which defines the position of the common ancestor of all nodes within the tree.

Dendrograms help to define the subtype of homology: orthology/paralogy

A dendrogram may reflect a species tree, but in general dendrograms are more complex, i.e. there are more leaves than species involved. One reason for this are gene duplications events in one or more species. Gene loss is possible as well, but is much less frequent. As mentioned in chapter 1.1, knowledge about the exact evolutionary relationship between proteins or genes is crucial for functional prediction. In this regard, dendrograms are an important means to define the two main subtypes of homology, which are orthology and paralogy.

Some practical considerations on dendrograms

Dendrograms are calculated from similarity data of protein or gene sequences involved. Similarity data can either be derived from pairwise comparisons as described above or from sequence features that make sequences differ from each other. Approaches relying on the latter try to find a gene tree with a minimum number of evolutionary events causing the observed tree topology at hand ('parsimony methods'). Another way to construct trees is to derive the distances between sequences or between groups of nearest neighbours. Here, a gene tree is constructed with the aim to minimize distances within the tree. This method implies an almost constant number of mutations within an evolutionary interval for the whole gene tree. The problem of different molecular clocks is solved more appropriately by stochastic models that estimate the maximum likelihood of a set of trees. For a detailed introduction to phylogenetic approaches see Durbin et al. and Saitou (Durbin, 1998, Saitou, 1987).

1.4 Data sources and functional prediction

1.4.1 Protein protein interactions and their prediction

Physical protein-protein interactions are fundamental to cell viability, growth, proliferation and many other biological processes within the living cell. Protein interactions take place between subunits of protein complexes, between distinct domains of mosaic proteins and in transient complexes between proteins that otherwise exist independently. During the last years, much attention has been paid to large-scale protein interaction analysis using different techniques like the two-hybrid system or the tandem-affinity purification technology (Gavin, 2002, Ho, 2002, Ito, 2001, Uetz, 2000). Several computational approaches exist in order to mine the rich interaction data sets or to derive protein interaction prediction methods. For example, systematic analyses of yeast two-hybrid data were performed by Sprinzak and Scheel (Scheel, 2001, Sprinzak, 2001). In both approaches, the occurrence of homology domains in pairs of interacting proteins were analysed for correlation. By statistical means, it could be shown that some combinations of homology domains were over-represented in the available interaction set. As a consequence, such combinations of homology domains could be used to infer interaction properties of other proteins containing these domains. Another approach predicts protein interactions from conserved gene based on the observation that genes with a conserved gene order across different species often encode proteins with mutual binding capacities (Dandekar, 1998). Marcotte and al. tried to infer protein interactions from the observation that two binding partners occasionally have homologues in another species, where they are fused into a composite gene (Marcotte, 1999). In a different approach, the analysis of protein-protein interfaces based on structural and sequence motifs has proven to be successful in the past (Jones, 1995). Here, general rules are derived from the interface regions of known protein complex structures and afterwards used to examine structures of test candidates for mutual interaction (Jones, 1996).

1.4.2 Model systems as source for biological data

The number of genes within the human genome is estimated to be in the range of 23.000 - 40.000 genes (Lander, 2001, Venter, 2001) and only a small portion of them has been characterized in detail. For the most part, functional annotations within genomic databases have been derived by sequence analysis as described above. However, computational prediction methods have their limits, for example in predicting cellular localization or cell-specific gene expression, and the need for complementary experimental approaches is obvious. One of the most widely used techniques to examine gene function is to analyse mutants. In this respect, the

mouse and its genome has become a mammalian model organism, which is more accessible to genetic modification than man for obvious reasons. At the same time, human and mouse share many anatomical, physiological and metabolic pathways making the mouse a useful model in genetic studies. However, even the mouse genome is far from being completely analysed in detail. Instead, many other model organisms, easier in use and more amenable to high-throughput approaches have gained importance in the recent past. For example, the zebrafish *Danio rerio* is considered as ideal for developmental studies due to its transparent embryos. Besides vertebrate model organisms, even simpler life forms like the nematode *Caenorhabditis elegans* with its body of just 959 well-defined cells or the popular unicellular eukaryote budding yeast (*Saccharomyces cerevisae)* can serve as models. In these organisms, fundamental cell processes can be studied and resulting information later be transferred to higher organisms like human. In this context, the homology approach plays a crucial role as pointed out in chapter 1.1.

Fundamental to using e.g. yeast as a model in the homology approach is the detection of homology between yeast and human genes despite their long evolutionary distance, but this has shown to be a feasible task (Foury, 1997). A connection between sequence similarity on one hand and functional similarity on the other, comes from the observation that mutational phenotypes in yeast can often be complemented by the ectopic expression of human cDNA clones encoding homologous genes (Schild, 1990). One of the main advantages of yeast as a model organism is its completely sequenced, small and compact genome with approximately 5800 to 6350 genes (Brachat, 2003, Goffeau, 1996). The compactness of the yeast genome is a result of the paucity of introns and a high gene density (~1 gene in 2 kb) (Goffeau et al., 1996). These advantages of yeast provided a basis for global studies that have focussed on functional analyses via systematic knockouts, protein localization, gene expression and protein interaction maps (Birrell, 2001, Gavin et al., 2002, Ho et al., 2002, Huh, 2003, Ideker, 2001, Ito et al., 2001, Ooi, 2001, Uetz et al., 2000, Winzeler, 1999). From these and previous studies, as well as from complementary experiments in human and other higher eukaryotes, the observations were made that yeast employs the same fundamental biological processes as higher eukaryotes, e.g. cell cycle, gene expression regulation, intracellular trafficking and protein degradation mechanisms. Therefore, yeast genetics provides a wealth of information, which to a certain degree also applies to the biology of a human cell.

Cliften et al. and Kellis et al. have shown that a comparative analysis of genomes worked successfully for the elucidation of genome evolution, the discovery of gene regulatory elements and the purpose of gene identification (Cliften, 2003, Kellis, 2004, Kellis, 2003). In these studies, genomes of closely related Hemiascomycetes with quite similar genome structures were

compared, and orthologous genes were identified according to their syntenic locations. Similarly, Boffelli et al. predicted functional regions in the genomes of primates (Boffelli, 2003). However, genome structures of yeast and human differ greatly, and synteny as a means for the assignment of orthologues is not appropriate. Often, comparable biological processes are more complex in human than in yeast, as can be seen from the expansion of gene sets underlying a certain cellular process. To draw conclusions on their evolution in yeast and human, sensitive comparisons of gene families using protein sequences and the profile technique as described above are more suited.

1.5 Aim of the study

The goal of this work was to analyse the ubiquitin-proteasome-pathway (UPS) and related pathways of *S. cerevisiae* and human by means of profile-based homology detection. This bioinformatical approach can be divided into three basic steps:

First, all proteins known to be relevant to the UPS should be catalogued both for human and *S. cerevisiae*. In this respect, the occurrence of homology domains, which are present at most if not all levels of the ubiquitylation pathway, made the profile technique a suitable means. The domain-specific profiles should be used to detect corresponding homologues in the proteomes of both species and to generate domain-specific catalogues of proteins.

Second, a gene tree should be inferred for each homology domain based on its respective multiple alignment. In this step, phylogenetic methods were used as described in chapter 2. The goal of this dendrogram analysis was to define orthologous protein pairs between *S. cerevisiae* and human for each UPS-relevant protein family. As additional approaches to the identification of orthologous pairs, the comparison of complete domain architectures and other methods of classical sequence comparison were used. In cases of a newly revealed evolutionary relationship between yeast and human proteins, existing knowledge on function should be tried to be transferred from one homologue to the other. The biological impact of each prediction should be analysed afterwards with a special focus on genes implicated in human diseases.

The goal of the third part of this work was to give insights into the evolution of the distinct subprocesses of the UPS system. Based on the results of the previous steps, the UPS repertoires of both species should be compared. For that purpose, each UPS-relevant protein family should be analysed for expansion and deletion events in the human and yeast genomes. Besides families, protein complexes with known or assumed importance for the UPS should be compared for their subunit composition.

2 Methods

2.1 Protein and nucleotide sequences for database searches

Protein sequence database searches for human were performed with a non-redundant data set in FASTA format constructed from current releases of UniProt (includes SwissProt, TrEMBL, PIR), GenPept and Ensembl peptides as described below. For generating this non-redundant data set, duplicate sequences found in multiple databases were discarded in the order SwissProt > TrEMBL > PIR > GenPept > Ensembl peptides with SwissProt being the preferred data source.

The UniProt protein database was obtained from the EBI (http://www.ebi.uniprot.org). A last update was performed in 02/05 with upgrading to UniProt release 4.0 (Bairoch, 2005). Even the non-redundant human portion of UniProt consisted of ~128,000 entries, which indicates that duplicates still existed as the size of the human genome is estimated to be in the range of 23.000 - 40.000 genes (Lander et al., 2001, Venter et al., 2001).

In addition to UniProt, GenPept release 146 (02/05) was included in the non-redundant database (ftp://ftp.ncifcrf.gov/pub/genpept/). Genpept is produced from translating coding regions of GenBank nucleotide sequences (Benton, 1990). GenPept was used in addition to UniProt, because of its higher update frequency.

Ensembl provides a polypeptide collection based on genomic sequence data (Hubbard, 2002). Thus, every entry has a corresponding genomic locus. The collection is subdivided into two parts, the 'Ensembl proteins' and the 'Ensembl GenScan predictions'. For Ensembl proteins usually experimental indications exist that these proteins are encoded in humans or other organisms. They are generated in an automatic procedure that regards existing proteins from human or other vertebrates, known mRNA or ESTs (Boguski, 1993, Curwen, 2004). In contrast, the peptide sequences belonging to the Ensembl GenScan group were predicted directly from the human genome using GenScan (Burge, 1997). GenScan-based transcripts and corresponding peptide sequences are identified in genomic DNA by an *ab initio* algorithm. Therefore, not every GenScan transcript and peptide sequence might reflect a proper protein.

Each protein sequence of interest for this work was tested for ESTs in order to examine its biological relevance. Ensembl entries were only taken into consideration, if no appropriate entry in the other, more reliable databases existed.

For several proteins of interest, only fragmentary polypeptide sequences were found in the protein databases. In these cases, full-length sequences were assembled from ESTs and genomic data provided by the NCBI and Ensembl project if possible.

The yeast protein sequences were downloaded from SGD (http://www.yeastgenome.org) (Issel-Tarver, 2002).

2.2 Databases for homology domain descriptors

A central point of this work was the classification of proteins into families based on homology domains and motifs. Several databases exist harbouring mathematical descriptors for homology domains, but only PROSITE and Pfam were used extensively in this work.

PROSITE (http://www.expasy.org/prosite/) consists of patterns and profiles of the 'generalized profiles' type that describe protein domains, motifs and enzyme families (Hofmann, 1999). The PROSITE entries are manually constructed and generally well curated. Only the profile section of PROSITE was used here, as patterns are not robust against deviations from the consensus pattern and do not take into account the conservation of the rest of the sequence they are matched to. These features make patterns much less sensitive than profiles. The PROSITE database distributes a software package to make use of the profile entries. This package, *pftools* 2.1, has been used here for database searching as well as for the construction of new profiles (see Table 2-1).

The Pfam database (http://www.sanger.ac.uk/Software/Pfam/) consists of profile Hidden Markov models (profile HMMs), which are generated from systematic analysis of domainwise MSAs of homology domains (Bateman, 2004). Like PROSITE, the Pfam database provides tools necessary to search protein sequences against a given profile HMM and to construct own HMMs (Eddy, 1998). Although the Pfam database is significantly larger than PROSITE, the set of homology domains relevant to the UPS are of the same size.

2.3 Constructing, refining and application of profiles

The construction of 'generalized profiles' can be subdivided into four major tasks, alignment generation, profile construction, profile scaling and iterative improvement. These tasks were carried out using different types of programs. An overview of this software and the corresponding WWW addresses of all programs used for profile construction are listed in Table 2-1.

Table 2-1 List of programs used described in this chapter.

program	package	URL
LALIGN	FASTA	ftp://ftp.virginia.edu/pub/fasta
ClustalW		ftp://ftp.ebi.ac.uk/pub/software/unix
T-Coffee		http://igs-server.cnrs-mrs.fr/~cnotred/
SPDB-Viewer		http://swissmodel.expasy.org/spdbv/
BLAST		ftp://ftp.ncbi.nih.gov/blast/
pfw pfmake pfscale autoscale psa2msa	pftools 2.1	http://www.isrec.isb-sib.ch/ftp-server/
readseq		ftp://ftp.bio.indiana.edu/molbio/readseq
dotter		ftp://ftp.sanger.ac.uk/pub/dotter
GDE		http://golgi.harvard.edu/ftp/

2.3.1 Obtaining an initial alignment

The first step in profile construction was to obtain a biologically reasonable alignment of complete sequences or subsequences suitable for profile construction. These sequences had to be homologous over their entire range or at least a local domain and were collected by BLAST searches (Altschul, 1990). Programs of choice for calculating the alignments were LALIGN for two-sequence local alignments and ClustalW as well as T-Coffee for MSAs (Chenna et al., 2003, Huang, 1991, Notredame et al., 2000). As ClustalW always creates global alignments, it is essential to exclude the non-homologous sequence stretches prior to MSA calculation. For that purpose, several not too divergent sequences were compared in a pairwise manner using a local alignment algorithm or by a dotplot analysis using the dotter program. If there were hints at multiple homology domains within one protein, e.g. if there were large non-homologous insertions between homologous regions, separate MSAs were calculated for each region.

For generating MSAs, fragmentary sequences are generally no problem and may be included in small numbers in the sequence set to be aligned. The standard gap penalties of ClustalW and T-Coffee appeared appropriate to keep the number of gaps at a low level. It is important to keep the number of gaps as small as possible in order to generate a profile that discriminates better against false positives, which is essential for the iterative profile improvement in later steps.

The generated MSAs were inspected and manually edited with GDE. The editing step included the removal of non-conserved residues at the domain borders, correction of alignment errors, minimizing the number of gaps and shifting of gap positions if necessary. In some cases, the alignment was adjusted according to a 3D structure available for one of the sequences in the alignment, i.e. gaps could be adjusted to secondary structure elements. If structures of at least two sequences were known, structural superpositions were calculated using the SPDB-Viewer

in order to generate a structurally correct alignment (Guex, 1997). Alignments were stored in MSF format for usage as starting point for following profile construction.

2.3.2 Construction of the profile

To generate a profile from an MSA in MSF format, weights for each sequence within the MSA were calculated using *pfw* from the *pftools* package. The reason for sequence weighting is to avoid a bias due to large subfamilies that might be part of the MSA. The *pfw* program performs a Monte-Carlo weighting for the MSA file with 2000 Monte-Carlo experiments chosen here.

Afterwards, the profile was constructed from the weighted MSA using the *pfmake* program, which is also part of the *pftools* package. The *pfmake* program was invoked with BLOSUM45 as substitution matrix and default penalties of 2.1 for gap opening and 0.2 for gap extension.

2.3.3 Scaling the profile and database search

When deciding whether a sequence belongs to a protein family, scores from searches with profiles in databases are only useful in combination with a score threshold separating random matches from true positive matches. To estimate what scores from a profile-sequence alignment may be considered significant, the profile has to be scaled. For that purpose, it is normally run against a randomised database (scaling database) lacking any true positive matches (Hofmann, 2000). From the scores of the matches, a score distribution can be plotted as depicted in Figure 2-1 A, in which intermediate scores constitute the main portion of resulted scores while high scores are rare and the frequency of high scores drops rapidly. In this high-scoring region, the score distribution follows an 'extreme value distribution', and can be plotted as a linear relationship between the log of the cumulative frequency and the score (Figure 2-1 B). Analysis of this function allows an extrapolation to the maximum expected score for a random match and thereby provides the possibility of defining a score threshold. This threshold is used for the generation of the scaling parameters, which are stored within the profile and allow to convert the raw score to an expectation value (E values).

Scaling of generalized profiles was carried out using the *autoscale* script from *pftools*. This program runs the raw profile against the scaling database (using *pfsearch*), invokes the proper scaling program *pfscale* to determine the scaling parameters and finally generates the scaled profile. The scaling database was built from randomisation of the SwissProt 34 database by individually inverting each protein sequence. The scores, and therefore the probability values

resulting from randomisation scaling of profiles, are much more reliable than those obtained from theoretical scaling approaches, which assume a uniform sequence composition (Hofmann, 2000). Scaled profiles were subsequently used for the database searches employing the *pfsearch* program from the *pftools* package.

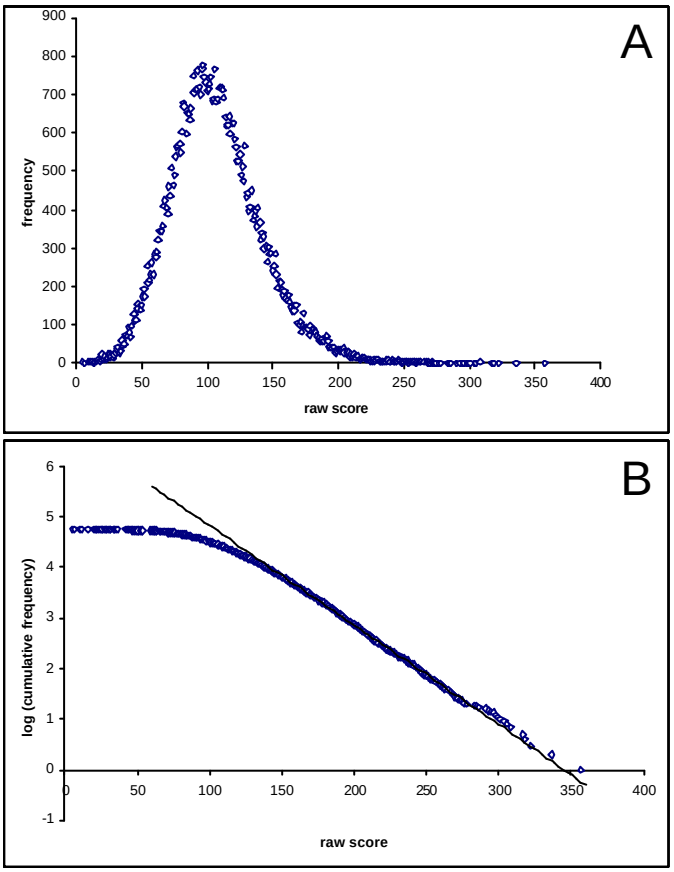


Figure 2-1 A) Raw score distribution of a profile constructed from PCI proteins against a randomised database. B) Decadic logarithm of the cumulative frequency plotted against the score. Approaching high scores, the function adopts the behaviour of an extreme value distribution and can be approximated by a linear function. Extrapolation of this function onto the abscissa yields the score expected from random matches.

2.3.4 Iterative improvement

The sensitivity and ability of a profile to discriminate between true and false protein family members can be enhanced remarkably by iterative refinement (see Figure 2-2) (Tatusov et al., 1994). First, the initial profile is run against a non-redundant database, which in this work was generated from all entries of UniProt and GenPept and therefore included various species. Besides the seed sequences present in the initial alignment, this database search was expected to identify additional proteins, which were not present in the initial alignment, but which represented homologues of the seed sequences. If the novel matches were judged true members of the domain family, they were added to the initial alignment and a new profile was constructed. Only sequence matches found with a probability of $P < 0.01$ were included into the subsequent round of profile construction. For that purpose, old and novel sequences were

aligned using the original profile as template ('profile-guided alignment'). The program used here was *pfsearch* with its '-x' option turned on. It runs the profile against the database, extracts the matching segments with scores above a given threshold and aligns them to the template defined by the profile. Afterwards, the output was converted by *psa2msa* and *readseq* to a multiple sequence alignment in MSF format suitable as input for *pfmake* in order to generate a new profile. MSAs were inspected and corrected manually prior to profile construction to eliminate alignment errors. This process of profile refinement was iterated several times until the results from profile searches converged, i.e. no novel sequence above threshold was found by the refined profile.

As a consequence of iterative improvement, the sensitivity of the profile will rise in general, i.e. further homologues of the integrated sequences can be expected in a following database search. At the same time, the profile's ability to discriminate between true and false positive matches will increase, because the higher the diversity of sequences in the multiple alignment, the clearer it becomes which positions are more conserved than others. The same holds true for gap positions or domain boundaries. A crucial point is the correctness of the multiple alignment and the absence of non-homologous sequences in the latter. The inclusion of non-homologous sequences inevitably leads to the detection and possibly inappropriate integration of even more non-homologous sequences resulting in mis-classifications.

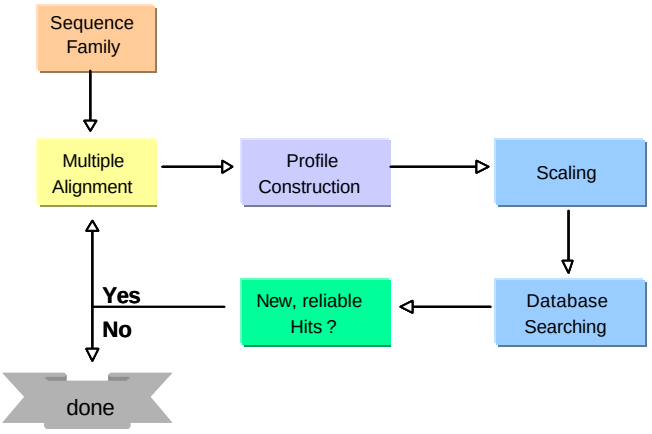


Figure 2-2 The concept of iterative profile improvement.

2.3.5 Determination of complete sets of protein families

In order to determine all the yeast and human members of a given protein family characterized by a homology domain, the corresponding profile was run against the species-specific datasets described in chapter 2.1. As the number of profiles for these searches was large and profile searches are computationally expensive, the results were stored in a permanent and easily accessible way. For that purpose, a simple relational database was set up, which stored all features returned from a successful profile match. These features are a unique protein identifier, the protein name, the name of the profile/homology domain, matching coordinates of the sequence, relative coordinates of the profile, etc.. For searches with generalized profiles, normalized scores and for searches with profile HMMs P-values were stored, respectively. For each profile match, one entry was added to the database.

For database implementation, the MySQL software was used. MySQL is a relational database system freely available from http://dev.mysql.com. It has several interfaces to current programming languages, but only the PERL interface, which was obtained from the MySQL site, was used in this work. Database maintenance and update procedure was performed by aid of PERL-based scripts.

To retrieve members of a given protein family, this database was queried using SQL, the standard language for accessing databases. After their retrieval, proteins were subjected to a proper non-redundancy check using an all-against-all-BLAST search strategy. The curated protein lists are shown in chapter 3. For visualization of the domain structure, a PERL-based script was developed using the GD-module and database queries as input.

2.4 Other tools

2.4.1 Dendrogram analysis

Dendrogram analysis was done with ClustalW, using the neighbour-joining algorithm (Saitou et al., 1987). Alignment columns containing gaps were removed prior to tree construction. The neighbour-joining trees were then tested by bootstrapping analysis with 1000 replicates. The programs treetool, which is part of the GDE package (Table 2-1), and Treeview were used for viewing and manipulating phylogenetic trees. Treeview can be obtained from http://taxonomy.zoology.gla.ac.uk/rod/treeview.html.

2.4.2 BOXSHADE

The alignments were shaded using BOXSHADE. It is available from (http://www.isrec.isb-sib.ch/ftp-server/).

2.4.3 Secondary structure prediction

Secondary predictions were obtained from sending the protein sequence alignments to the Jpred and PHD servers (http://cubic.bioc.columbia.edu/predictprotein/) (Cuff, 1998, Rost, 2003). Jpred predictions rely on a neural network strategy and assign either α-helix, β-sheet or random coil states to the positions of the provided alignment. The principle of how PHD generates its predictions is similar.

2.4.4 18S rRNA tree

For analysing the phylogenetic distribution of the SUMO interaction motif (SIM) described in chapter 3.6.7, an 18S rRNA based tree was constructed for selected species. The 18S rRNA gene sequences were aligned by using ClustalW. The alignments were adjusted manually. Phylogenetic analyses were performed also by using ClustalW and an unrooted phylogenetic tree was constructed by using the neighbour-joining method. The stability of the individual branches was assessed by using the built-in bootstrap method of the ClustalW program. By aid of PERL scripts the phylogenetic profile of a SIM for a given protein was mapped onto the 18S rRNA tree for visual inspection.

3 Results

3.1 Ubiquitin and its relatives

3.1.1 Type I Ubiquitin-like modifiers

3.1.1.1 Ubiquitin in the yeast and human genome

Ubiquitin (Ub) is a highly abundant protein in the eukaryotic cell. Its expression is unusual, as it is encoded as a poly-ubiquitin precursor or fused to the ribosomal proteins L40 and S27a, which are normally highly expressed as well (Ozkaynak, 1987, Ozkaynak, 1984, Redman, 1989). Prior to the usage of the Ub moieties contained in these polypeptides, the precursors have to be processed, i.e. the fusion proteins have to be cleaved. A role of Ub fused to ribosomal proteins in ribosome biogenesis has also been suggested (Finley et al., 1989).

Screening genomic sequence databases, four distinct ORFs coding for Ub in both yeast and human were detected (see Table 3-1). In yeast, three of these ORFs consist of one Ub moiety fused to an L40 or S27a ribosomal protein and one ORF encodes a poly-Ub containing 5 copies. In comparison, the human genome contains two poly-ubiquitin genes, UBB and UBC, which harbour 3 or 9 Ub moieties, respectively. Two additional Ub encoding genes have their Ub moiety fused to either L40 or S27a. Therefore, yeast and human store their Ub-coding regions in a similar manner. The number of pseudogenes derived from human Ub-coding genes is remarkable (see Table 3-2). At least nine different Ub-pseudogenes were detected, six of which seem to have originated from RPS27A. The remaining four are likely processed pseudogenes of UBB and UBA52. In nearly all cases, the Ub-pseudogenes have arisen from retrotransposition (data not shown).

Table 3-1 Ub-like modifiers in yeast and human. Column 'CT' shows the amino acids homologous to the terminal 'GG' motif in Ub. Column ‘orphan’ indicates if orthologues are present in human or yeast, respectively.

	Ub-like modifier Gene name	ORF/Uniprot	CT	without orthologue	process
Yeast	Ub: UBI4	YLL039C	GG	-	UPS, protein sorting
	Ub: RPL40A	YIL148W	GG	-	UPS, protein sorting
	Ub: RPL40B	YKR094C	GG	-	UPS, protein sorting
	Ub: RPS31	YLR167W	GG	-	UPS, protein sorting
	RUB1	YDR139C	GG	-	cullin/RING-E3 regulation
	SMT3	YDR510W	GG	-	nuclear transport, localization
	URM1	YIL008W	GG	-	stress, invasive growth
	ATG8	YBL078C	FG	-	autophagy, CVT pathway
	ATG12	YBR217W	FG	-	autophagy, CVT pathway
	HUB1	YNR032C-A	YL	-	polarized morphogenesis
Human	Ub: UBA52	13569612	GG	-	UPS, protein sorting
	Ub: RPS27A	Q5RKT7	GG	-	UPS, protein sorting
	Ub: UBB	Q5U5U6	GG	-	UPS, protein sorting
	Ub: UBC	Q7L684	GG	-	UPS, protein sorting
	NEDD8	Q15843	GG	-	cullin/RING-E3 regulation
	SMT3A/SUMO-3	P55854	GG	yes	conjugated under stress
	SMT3B/SUMO-2	P61956	GG	yes	conjugated under stress
	SMT3C/UBL1/SUMO-1	P63165	GG	-	nuclear transport, localization
	SUMO-4	Q6EEV6	GG	yes	heat shock induced
	SUMO-5/UBL6	Q6P094	GG	-	?
	UBL3	O95164	ET	yes	?
	UBL4	P11441	EK	yes	?
	UBL5	Q9BZL1	YQ	-	polarized morphogenesis
	GABARAP/FLC3B	O95166	YG	-	autophagy, CVT pathway
	GABARAPL1/GEC1	Q9H0R8	YG	-	?
	GABARAPL2	P60520	FG	-	intra-Golgi traffic
	GABARAPL3	Q9BY60	YG	-	?
	APG12L	O94817	WG	-	autophagy, CVT pathway
	Fat10	O15205	GG	yes	apoptosis, interferon response
	ISG15/UCRP	P05161	GG	yes	immune response
	FAU/FUBI/MNSF	P35544	GG	yes	T-cell activation
	Ufm1	P61960	VG	yes	regulation
	MOCS2	O96033	GG	yes	molybdoterin synthesis
	C9orf74	Q9BTM9	GG	-	stress (by similarity)

Table 3-2 Predicted human pseudogenes of Ub and related modifiers. The identification of pseudogenes is described in chapter 2. Accession numbers are Uniprot for Q6ZRT8 and Q9BX44, otherwise EnsEMBL peptide IDs or Genscan IDs are given. For SUMO-1-like 1 only chromosomal coordinates were available.

Gene name	Accession number	Chromosome
Ub: RPS27A-like 1/bA92K2.2	Q9BX44	Chr1
Ub: RPS27A-like 2	GENSCAN00000014306H	ChrX
Ub: RPS27A-like 3	GENSCAN00000048511H	Chr6
Ub: RPS27A-like 4	ENSP00000355184	Chr2
Ub: RPS27A-like 5	ENSP00000334842	ChrX
Ub: UBA52-like 1	ENSP00000258728	Chr7
Ub: UBA52-like 2	GENSCAN00000058681H	Chr9
Ub: UBB-like 1	ENSP00000320067	Chr2
Ub: UBB-like 2	Q6ZRT8	Chr17
FAU-like	ENSP00000335590	Chr18
FAU-like2	ENSP00000310146	Chr11
SUMO-1-like 1	Chr1, base 157100317:157101631	Chr1
SUMO-2-like 1	ENSP00000313744	ChrX
SUMO-2-like 2	ENSP00000328831	Chr7
SUMO-2-like 3	GENSCAN00000025561H	Chr8

3.1.1.2 Type I Ubiquitin-like modifiers

Among the 20 human and six yeast type I Ub-like modifiers, only few have been analyzed in detail like human NEDD8 or SUMO1 as well as their yeast orthologues. At least the broad biological process of most type I Ub-like modifiers is known (Table 3-1).

The human genome encodes five SUMO paralogues while yeast has only one SUMO protein encoded (Smt3) (see Table 3-1). Four of the five human SUMO proteins have been described so far and functional data is available (Dohmen, 2004). The fifth one is currently only mentioned in the Vertebrate Genome Annotation (VEGA) database and the name UBL6 has been suggested (Ashurst, 2005). From sequence analysis done here, UBL6 has a clear evolutionary connection to the SUMO subfamily of type I Ub-like modifiers and is referred to as SUMO5 in this work. SUMO5 shares 87% sequence similarity with SUMO1 and is therefore more closely related to the classic SUMO than SUMO2-4. It maps to chromosome 20 and appears to be weakly expressed (only in testis) compared to SUMO1 (Strausberg, 2002). However, any functional characterization of the SUMO5 protein is still unavailable. As introns are missing and the 3'-UTR of SUMO5 is related to the 3'-UTR of SUMO1, the SUMO5 gene probably has arisen from retro-transposition of SUMO1 mRNA. A 'GG'-motif is present at the C-terminus of the gene product as is the case for SUMO1-4. In addition to these five expressed SUMO genes, at least four SUMO pseudogenes exist in the human genome, which probably have arisen from SUMO1 and SUMO2 mRNA retrotransposition (see Table 3-2).

For yeast Rub1 and Urm1, clear human orthologues could be proposed, which are NEDD8 and C9orf74, respectively (see Table 3-3). Human Ufm1 is absent from yeast and other fungi. Another class of type I modifier, which is completely missing in yeast, are linear diubiquitin modifiers like Fat10 or ISG15. As the latter carries out a function in the innate immune system (Kim et al., 2003), which is specific to multicellular organisms, it is not expected to be found in unicellular organisms like yeast.

The Ub-like protein FUBI (35.5% identity to Ub) is encoded by the human gene FAU, in which it is fused to a ribosomal subunit (S30). FAU is not found in the yeast genome. Like the Ub-precursors, the initial FAU gene product is post-translationally cleaved in order to allow S30 incorporation into the small ribosomal subunit. FUBI has been reported to get covalently linked to a murine Bcl2-like protein, which is orthologous to human pro-apoptotic BCL-G (Nakamura, 2003). Within the human genome, two paralogues of the FAU gene have been detected, which are likely to be pseudogenes. FAU-like on chromosome 18 seems to be the result of the retrotransposition of the original FAU-mRNA, as no intron exists and a polyA-tail is detectable

in the genomic sequence, ~15 bases downstream of the 'AATAAA'-motif. In agreement with this finding, FAU-like has been described as pseudogene FAUP1 elsewhere (Kas, 1995). A second putative FAU pseudogene, here named FAU-like2, was found on chromosome 11. This gene has the same intron as FAU, but lacks the start-codon.

Three sequences (UBL3, UBL4, UBL5) were found in the human genome that share a high similarity to the original Ub sequence. All three genes are expressed. However, a role as modifier seems questionable, as these three proteins lack the C-terminal 'GG' motif thought to be characteristic for type I modifiers. The yeast orthologue of UBL5 is Hub1, which has been reported to form adducts with other proteins, though it is not clear, whether these adducts are covalent or noncovalent in nature (Wilkinson, 2004). In contradiction to Wilkinson et al., Yashiroda et al. have ruled out a function of Hub1 from *S. pombe* as a classic type I modifier (Yashiroda, 2004).

Table 3-3 Orthology assignments according to sequence comparisons in a dataset of Ub-like type I modifiers.

Yeast	Human
Ubiquitin	Ubiquitin
RUB1	NEDD8
SMT3	(SMT3C/UBL1/SUMO-1), (SUMO-5/UBL6)
HUB1	UBL5
ATG8	GABARAP, GABARAPL1-3
ATG12	APG12L
URM1	C9orf74

3.1.2 Type II Ub-like proteins

3.1.2.1 Ub-like domains detected by the Ub-profile

Besides type I Ub-like proteins, a set of multi-domain proteins with an embedded Ub-like domain was found in yeast and human (see Table 3-4). These proteins were detected by the same Ub-based profile that was also used for the detection of most of the type I Ub-like proteins. Often, these so called type II Ub-like proteins have their Ub-like domain positioned at the extreme N-terminus (see Figure 3-1). They lack the 'GG' motif at the C-terminus of the Ub-like domain (data not shown). Therefore, it is likely that type II Ub-like proteins cannot be processed and conjugated (Jentsch et al., 2000).

The human Ub-like type II protein set consists of 59 members while the yeast set contains only twelve proteins. Prominent members are yeast Rad23 and Dsk2 as well as their human orthologues (see Table 3-4), which all have been studied extensively (Luders et al., 2003, Walters, 2002). These proteins possess one or two UBA domains in addition to their Ub-like domain (see Figure 3-1).

Interestingly, Ub-like domains are found in many proteins with a direct connection to the UPS like DUBs, E1s and substrate delivery factors like Rad23. Moreover, Ub-like domains could be detected in two subunits of the IKK complex, CHUK/IKKA and IKBKB, responsible for phosphorylation of NF-κB inhibitors from the IκB family. Such phosphorylation immediately triggers the rapid Ub-mediated proteasomal proteolysis of the IκB inhibitor, thereby restoring NF-κB function as a transcription factor. The observation of Ub-like domains in these both NF-κB related kinases as well as in another protein involved in the NF-κB pathway, IKBKE, is probably tightly linked to the Ub-dependent degradation of the inhibitors. Moreover, this example shows that the occurrence of typically UPS-associated homology domains in seemingly unrelated cellular processes, here NF-κB signalling, may indicate the involvement of the UPS on a regulatory level in these processes.

Table 3-4 Ub-like type II proteins. Additional domains were provided as well if possible. Orthologues are defined in the upper panel with different assignments separated by grey/white transitions. Bold letters denote proteins without Ub-like domain whose orthologues were found to contain one. Pseudogenes are printed in italics.

Ub-like type II proteins					
Yeast		**additional**	**Human**		**additional**
Gene name	**ORF**	**domains**	**Gene name**	**Uniprot number**	**domains**
DDI1	YER143W	UBA, Asp peptidase	DDI1	Q8WTS3	UBA, peptidase
			DDI2	Q7RTZ0	UBA, peptidase
PAC2	YER007W	LRR	TBCE	Q15813	LRR
NPL4	YBR170C		NPL4	Q8TAT6	NZF_RANBP
YOD1	YFL044C	OTU	FLJ46133	Q6ZRS6	OTU
DSK2	YMR276W	UBA	UBQLN1/PLIC1	Q9UMX0	UBA
			UBQLN2/PLIC2	Q9UHD9	UBA
			UBQLN3	Q9H347	UBA
			UBQLN4/UBIN	Q9NRR5	UBA
RAD23	YEL037C	UBA	RAD23A	P54725	UBA
			RAD23B	P54727	UBA
UBA1	YKL210W	UBA_NAD	UBE1	P22314	UBA_NAD
UBA2	YDR390C	UBA_NAD	UBLE1B/UBA2/SAE2	Q9UBT2	UBA_NAD
UBP6	YFR010W	USP	USP14	P54578	USP
UBA3	**YPR066W**	**no Ub-like domain**	UBE1C/UBA3	Q8TBC4	UBA_NAD
PRP21	**YJL203W**	**no Ub-like domain**	SF3A1	Q15459	SURP
Yeast members without orthologues					
Gene name	**ORF**	**domains**			
ESC2	YDR363W				
USA1	YML029W				
YOL111C	YOL111C				
Human members without orthologues					
Gene name	**Uniprot number**	**domains**	**Gene name**	**Uniprot number**	**domains**
BAG1	Q99933	BAG	USP31	Q86UV5	USP
GABPA	Q06546	ETS	USP32	Q8NFA0	USP
FBXO7	Q9Y3I1	Fbox	SACS/Sacsin	Q5T9J7	DNAJ_N, HEPN
CHUK/IKKA	O15111	kinase	A-735G6.2	O14562	
IKBKB	O14920	kinase	ANUBL1	Q86XD8	
IKBKE	Q14164	kinase	BAT3	P46379	
FLJ46103	Q6ZRU1	Ttrap_NT	C16orf33	Q9BV90	
MGC10067	Q8WVY7	NIF	C7orf21	Q9BVT8	
VCIP135	Q96JH7	OTU	DC-UbP	Q8WUN7	
PARKIN	O60260	Parkin triad (RING)	FAFX	Q93008	
UBCE7IP3/XAP4	Q9BYM8	Parkin triad (RING)	FAFY	O00507	
UHRF1/NP95	Q96T88	RING	FLJ22313	Q9BSE4	
UHRF2/NIRF	Q96PU4	RING	FLJ35834	Q8NA54	
Catastrophin	Q5QJ74	UBA	FLJ90280	Q8NCF5	
NUB1/NYREN18	Q9Y5A7	UBA	HERPUD1/MIF1	Q15011	
RAD23C	*ENSP00000334233*	*UBA*	KIAA0633	O75128	
SB132/BMSCUBP	Q96S82	UBA	LOC164153	Q8N7F7	
UBADC1/GBDR1	Q9BSL1	UBA	novel-UBL-type-II 1	GENSCAN00000058079H	
mop-4	Q9H3T7	UBA_NAD	OASL	Q15646	
UBE1L/UBE2	P41226	UBA_NAD	TCEB2/Elongin B	Q15370	
FAF1	Q9UNN5	UBX	*TCEB2/Elongin B like*	*ENSP00000333957*	
USP24	Q9UPU5	USP	UBTD1	Q9HAC8	
			UNQ1897	Q71RG4	

Two of the Ub-like type II proteins from human have a 'GG' motif that is homologous to the 'GG' in Ub-like type I modifiers. While in ANUBL1 the Ub-like domain is localized at the N-terminus with much primary sequence following, the homologous 'GG' of SF3A1 is only three residues away from the C-terminus. SF3A1 is a well known subunit of the spliceosome but a processing similar to that found in several Ub-like type I modifiers has not been described for SF3A1 (Das, 2000). Remarkably, the yeast orthologue Prp21 lacks the Ub-like domain (Table 3-4). This observation probably indicates that the Ub-like domain is dispensable for splicing related functions of Prp21/SF3A1 in these species.

Interestingly, FAF1, an enhancer of FAS-induced apoptosis has both the Ub-like and a Ubx domain merged into one polypeptide (see Figure 3-1) (Ryu, 1999). The Ubx domain has been reported to adopt a Ub-like fold (Buchberger et al., 2001). However, only in a few cases, Ubx based profiles were able to detect proteins with a proper Ub-like domain and vice versa (data not shown). Apart from the weak sequence similarity, proteins with Ubx or Ub-like domain are frequently found together with other homology domains typically associated with the UPS, e.g. UBA domains (see Figure 3-1).

3.1.2.2 Remotely Ub-like domains in Ub-activating enzymes

The structures of the human Ub-activating enzymes, UBE1C/UBA3 and UBLE1B/UBA2, clearly demonstrate the presence of a Ub-fold domain outside the catalytic domain (Lois et al., 2005, Walden et al., 2003). In most E1s, this Ub-fold domain is too divergent to be matched by the classic Ub-based profile. Therefore, new profiles were generated based on the Ub-fold domain of the crystallized human E1s for an in-depth analysis of the Ub-fold domain in other proteins of the E1 family.

By means of a rigid-body superposition of the Ub-fold domains in human UBE1C/UBA2 and UBLE1B/UBA3, a structural alignment was derived and converted to a profile, which afterwards was run against E1 sequences. As a result, three human E1s as well as many other E1s from different species were found to contain a Ub-fold domain (see Table 3-4 and Figure 3-2 for a domain structure). These human proteins are the Ub-activating UBE1 and the ISG15-activating UBE1L as well as mop-4, which has no known modifier to activate so far. Remarkably, the Ub-fold domain is absent from yeast Uba3 and its fungal orthologues.

Figure 3-1 Domain structure of selected proteins with Ub-like domains. The Ub-like domain is coloured in green. A frame indicates orthologues from human ('hs') and yeast ('sc'). For yeast Dsk2, only one of the four orthologous Ubiquilins is shown, UBQLN1.

3.2 Activating enzymes for Ub and related modifiers (E1)

3.2.1 E1 protein sets for yeast and human

The UBA_NAD profile of PROSITE covers the NAD-binding region of typical E1s. By means of this profile, several established and putative E1s were identified (see Table 3-5). The largest group of E1 enzymes harbours two copies of the NAD-binding region in one polypeptide, for example the Ub-activating Uba1 (yeast) or UBE1 (human). By contrast, the SUMO- and NEDD8-activating enzymes act as heterodimers, whose subunits carry only one UBA_NAD copy each. The homodimeric yeast Atg7 and its human orthologue APG7L as well as the likely homodimeric Uba4 (a self-interactor in yeast) also contain only one NAD-binding region.

As the primary sequences of the most familiar E1s are well conserved between yeast and human, clear 1:1-type orthologous pairs could be defined. Concomitantly, the domain topology is preserved for each pair (see Figure 3-2). They are listed in Table 3-6 together with their substrates and corresponding E2s.

Table 3-5 Activating enzymes from yeast and human. Gene names are shown together with systematic ORF names and protein accession numbers, respectively. Orthologues are opposed in the same row and different pairs are separated by grey/white transitions.

Yeast E1 Gene name	ORF	Human E1 Gene name	Uniprot number
UBA1	YKL210W	UBE1	P22314
UBA2	YDR390C	UBLE1B/UBA2/SAE2	Q9UBT2
UBA3	YPR066W	UBE1C/UBA3	Q8TBC4
UBA4	YHR111W	MOCS3	O95396
ULA1	YPL003W	APPBP1	Q13564
AOS1	YPR180W	UBLE1A/AOS1/SAE1/SUA1	Q9UBE0
ATG7	YHR171W	APG7L	O95352
YHR003C	YHR003C	-	-
YKL027W	YKL027W	-	-
-	-	UBE1L/UBE2	P41226
-	-	UBE1DC1/Uba5	Q9GZZ9
-	-	mop-4	Q9H3T7

Table 3-6 Known Ub-like modifiers and their activating and conjugating enzymes in yeast and human. Atg12, Atg8 and Urm1 are distinct from the other modifiers because they are unrelated in sequence to Ub. Whether Hub1 functions as a modifier is currently unclear.

Yeast Modifier	**E1**	**E2**
Ub	Uba1	Ubc1-8, Ubc10-11, Ubc13
Smt3	Aos1-Uba2	Ubc9
Rub1	Ula1-Uba3	Ubc12
Urm1	Uba4(-Uba4?)	-
Atg8	Atg7	Atg3
Atg12	Atg7	Atg10
Hub1	?	?
Human Modifier	**E1**	**E2**
Ub	UBE1	many, see chapter 3.3
SUMO-1	UBLE1A/SAE1-UBLE1B/SAE2	UBE2I
NEDD8/Rub1	APPBP1-UBE1C/UBA3	UBE2M
Urm1/C9orf74	MOCS3	-
GABARAP (Atg8-like)	APG7L	APG3L
APG12L	APG7L	APG10L
ISG15	UBE1L/UBE2	UBE2E2/UBCH8
Ufm1	Uba5	Ufc1/HSPC155
Fat10	?	?
FUBI/FAU	?	?
UBL5	?	?

3.2.2 Specific genes and homology domains in the E1 family

In addition to a common set of E1 enzymes found in both species, both yeast and human have specific E1 copies that occur only in that particular species. For two of these human proteins, UBE1L/UBE2 and Uba5, the Ub-like protein to be activated is known, which is ISG15 and Ufm1, respectively (Komatsu et al., 2004, Yuan, 2001). The ISG15 activator UBE1L/UBE2 is closely related to the Ub-activator UBE1, and has another close paralogue with an identical domain topology, mop-4. The latter protein is highly expressed in monocytes, but its substrate is unknown (Takayama 1998, unpublished).

In addition to the known activators of the known Ub-like proteins, yeast also contains two orphan proteins with an E1 architecture (Yhr003c and Ykl027w). Both genes seem to have arisen from gene duplication and share only weak sequence similarity with Uba1. Interestingly, their C-terminus is probably homologous to the human hypothetical protein FLJ36074 (Q8N9Y2), which in turn lacks the UBA_NAD domain, the hallmark of E1 enzymes. While these two ORFs are specific to yeast, it is remarkably that the yeast genome does not encode any orphan type I Ub-like modifiers. It is currently not clear if the corresponding proteins have a redundant function in activating type I Ub-like modifiers or a E1 function at all (see chapter 4.2.1 for a more detailed discussion).

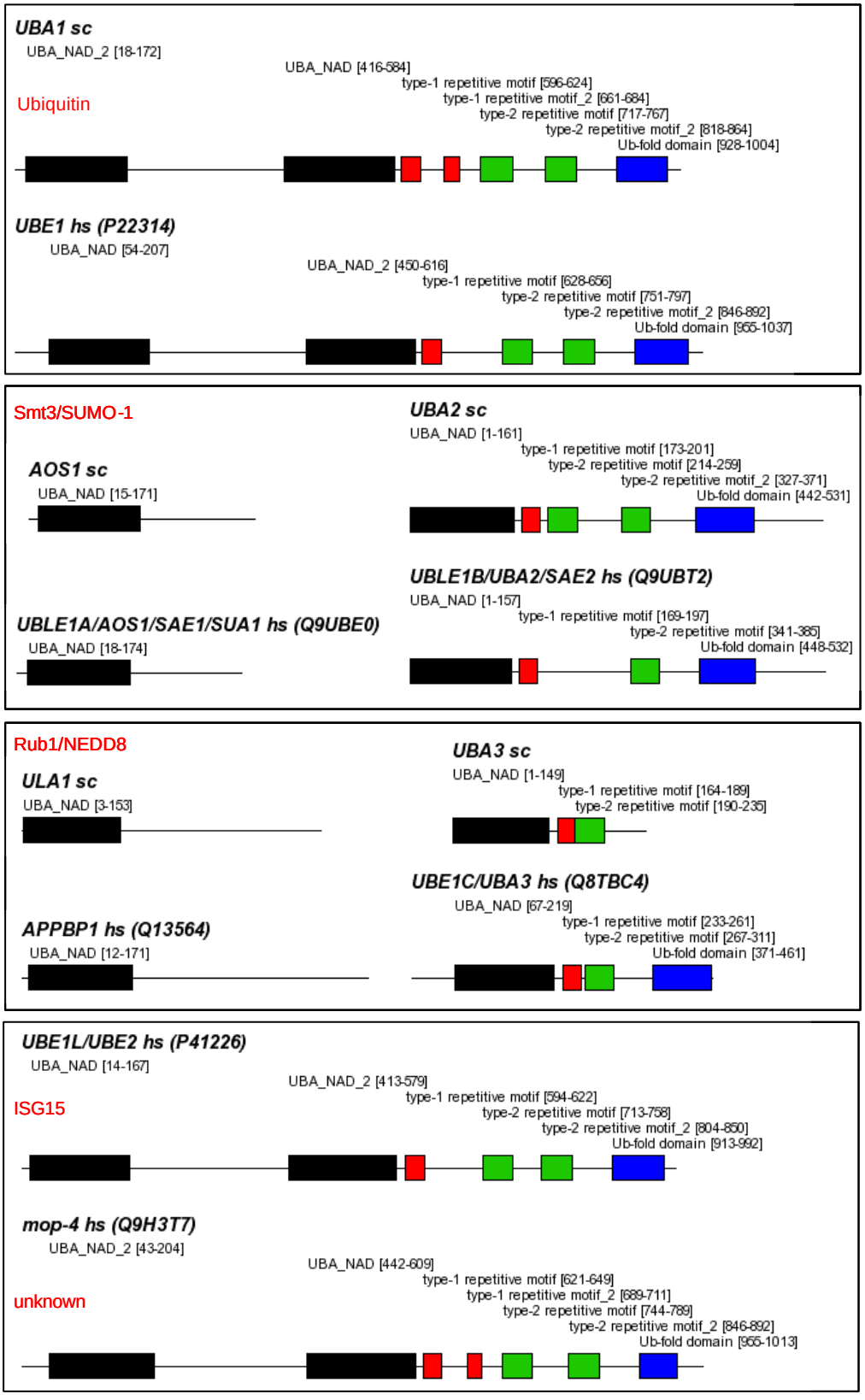


Figure 3-2 Domain structure of yeast and human E1 enzymes with a similar C-terminal domain arrangement in the active subunits/monomers. Each box contains E1 enzymes responsible for the activation of the modifier given in red letters. As far as possible, yeast and human orthologues are opposed indicated by the species abbreviations ‚sc' and ‚hs' in each description line, respectively. Two different proteins within a row reflect the subunits of heterodimeric E1s. The type-I repetitive motif that follows a UBA_NAD domain contains the catalytic cysteine.

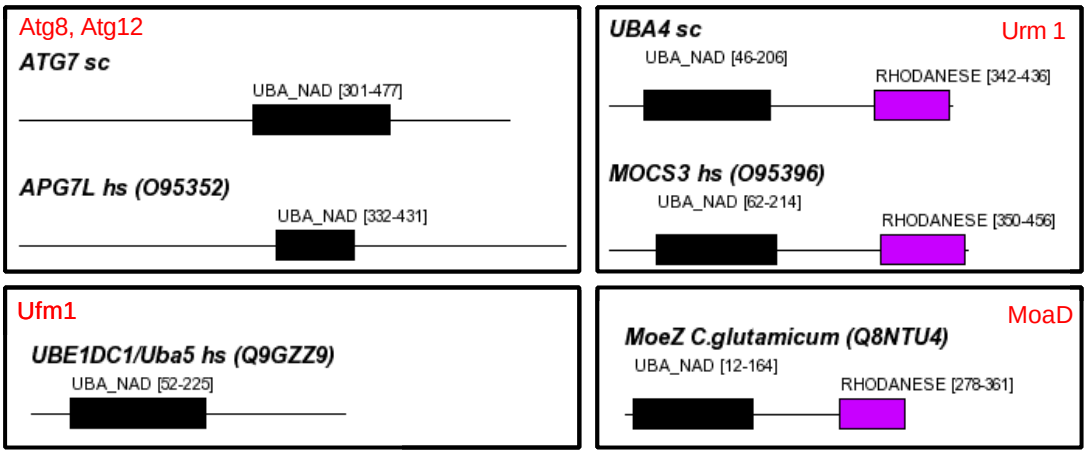


Figure 3-3 Domain structure of E1s for Atg8/Atg12, Urm1, Ufm1 and the MoaD from *C. glutamicum*, MoeZ. The catalytic cysteine in E1s containing a Rhodanese domain is located few residues downstream of the UBA_NAD domain (not shown).

3.2.3 Repetitive motifs in some E1s contain the active site

Two new and distinct repetitive motifs (type-1/type-2) were found downstream of the UBA_NAD domain in E1s (see Figure 3-2). The relative position of these motifs is quite variable. In general, the type-1 repetitive motif, which directly follows the UBA_NAD domain, harbours the catalytic cysteine. A second copy located downstream is devoid of a cysteine at the corresponding position. The presence of the type-2 repetitive motif does not seem to be obligatory, as it could not be detected in all E1s so far. A more general observation is that type-1 motifs always precede the type-2 motif(s). The insertion length between the two different repeat stretches is highly variable, e.g. they are immediately adjacent in UBE1C/Uba3, while there are large insertions for Uba2/SAE2. There seems to be a slight preference for two copies of the type-2 motif in E1s with two UBA_NAD domains, i.e. in E1s, which are active as monomers.

3.2.4 Other domain arrangements in E1

Not all E1s share the same domain arrangement as found in yeast Uba1 and other E1s as shown in Figure 3-2. Uba4 and MOCS3 have a Rhodanese domain C-terminally of their UBA_NAD domain (see Figure 3-3). This domain arrangement is also detected in E1-like bacterial proteins mediating MoaD and ThiS activation, which is specifically adapted to the function of these two structural homologues of Ub and therefore not found in other E1s (Rajagopalan, 1997). Atg7/APG7L and Uba5 neither share the Rhodanese domain nor show a domain arrangement of repetitive motifs and a Ub-like domain. Probably, their function does not rely on auxiliary domains in addition to the UBA_NAD or such domains have not been discovered so far.

3.3 Ub-conjugating enzymes (E2)

3.3.1 E2 protein sets for yeast and human

To find sequences related to the E2s known so far, a profile adapted from the Ubc homology domain was used for database searches. Overall, 15 proteins from yeast and 42 from human were detected with significance. These sequences can be divided into two groups based on the presence or absence of the catalytic cysteine. Proteins with a catalytic cysteine are considered active E2s, while proteins lacking that special cysteine are thought to be inactive Ubc enzyme variants (UEV) (Ponting et al., 1997). As the Ubc domain in UEV proteins has been reported to have Ub-binding properties, UEV proteins are described in chapter 3.6. Therefore, Table 3-7 covers just active Ubc proteins including autophagy E2s.

For almost all yeast E2 proteins with a Ubc domain, human orthologues could be defined (see Table 3-7). In some instances, the human ancestor gene seems to have been duplicated leading to 1:n-relationships, but in the case of yeast Ubc4 and Ubc5 the fungal organism shows also evidence of gene duplication after its branching off from the last common ancestor of yeast and human. Only the yeast peroxisome biogenesis factor Pex4 could not be assigned to a human orthologue. As Pex4 almost exclusively consists of the Ubc domain, regions outside this domain did not help in finding a human orthologue. While there is only one yeast-specific Ubc protein, the human genome has a considerably higher number of unique Ubc genes, including the conjugating enzyme for the human-specific modifier ISG15 (Zhao et al., 2004).

Table 3-7 List of yeast and human E2s. Orthology assignments are given in a subtable. For autophagy E2s, orthologous pairs have directly been opposed.

Yeast Ubc-E2	
Gene name	**ORF**
CDC34	YDR054C
PEX4	YGR133W
QRI8	YMR022W
RAD6	YGL058W
UBC1	YDR177W
UBC11	YOR339C
UBC12	YLR306W
UBC13	YDR092W
UBC4	YBR082C
UBC5	YDR059C
UBC6	YER100W
UBC8	YEL012W
UBC9	YDL064W

Assignment orthologues	
Yeast	**Human**
CDC34	CDC34, UBC3B
QRI8	UBE2G2
RAD6	UBE2B, UBE2A
UBC1	HIP2
UBC11	UBE2C
UBC12	UBE2M
UBC13	UBE2N
UBC4, UBC5	UBE2D1, UBE2D2, UBE2D3, UBE2D4, OTTHUMP00000030191
UBC6	UBE2J2
UBC8	UBE2H
UBC9	UBE2I

Human Ufm1-E2	
Gene name	**Uniprot number**
Ufc1/HSPC155	Q9Y3C8

Human Ubc-E2	
Gene name	**Accession number**
BIRC6	Q9NR09
CDC34	P49427
FLJ11011	Q96B02
FLJ13855	Q9H832
HIP2	P61086
HSPC150	Q9NPD8
KIAA1734	Q9C0C9
LOC92912	Q8WVN8
MGC42638	Q8IWF7
NCE2	Q969M7
UBE2A	P49459
UBE2B	P63146
UBE2C	O00762
UBE2D1	P51668
UBE2D2	P62837
UBE2D3	P61077
UBE2D4	Q9Y2X8
UBE2E1	P51965
UBE2E2/UBCH8	Q96LR5
UBE2E3	Q969T4
UBE2G1	P62253
UBE2G2	P60604
UBE2H	P62256
UBE2I	P63279
UBE2J1	Q9Y385
UBE2J2	Q8N2K1
UBE2L3/UBCH7	P68036
UBE2L3-C13	Q5VZ96
UBE2L3-C14	10444495
UBE2L6	O14933
UBE2M	P61081
UBE2N	P61088
UBE2Q	Q7Z7E8
UBE2QL	ENSP00000339662
UBE2R2	Q712K3
UBE2S	Q16763
UBE2SL	Q6NXQ4
UBE2U	Q8N1D4

Yeast autophagy E2s	
Gene name	**ORF**
ATG3	YNR007C
ATG10	YLL042C

Human autophagy E2s	
Gene name	**Uniprot number**
APG3L	Q9NT62
APG10L	Q6PIX1

3.4 Ligases for Ub and related modifiers

3.4.1 Finding RING finger proteins

One class of Ub ligase is characterized by the presence of a complex Zn finger domain termed 'RING finger' and acts by binding to Ubiquitin conjugating enzymes bringing them into close contact with the substrate. Several sequence-based approaches for the classification of complex Zn fingers are currently in use. The simplest and most accessible one relies solely on the nature and spacing of the cysteine and histidine residues that act as ligands for the Zn(II) ion. However, a number of RING finger proteins deviates from the consensus Cys/His pattern (Aravind et al., 2000). Therefore, profiles are a more exact method to describe the RING finger domain, as the profile technique allows also to consider important residues outside the zinc-binding residues.

3.4.1.1 Construction of a pure RING finger profile

In a first approach, members of the RING finger family were identified by means of the RING finger profile entry of PROSITE. There was a markedly high portion of sequences that tested positive for both the RING finger and a second, functionally distinct type of zinc finger, the PHD finger. This observation was the reason to construct 'pure' profiles for each the RING and the PHD finger family in order to enable a clearer classification of novel proteins. To gain such pure profiles, only proteins with an unambiguous PHD/RING status were used in the initial step of profile generation. As a consequence, the resulting profiles had a slightly reduced sensitivity but yielded a clearer separation between RING and PHD domains. As evident from Figure 3-4, RING and PHD finger proteins form distinct clusters in the two-dimensional score space. Therefore, the newly constructed RING finger profile was employed to retrieve true RING finger members in yeast and human.

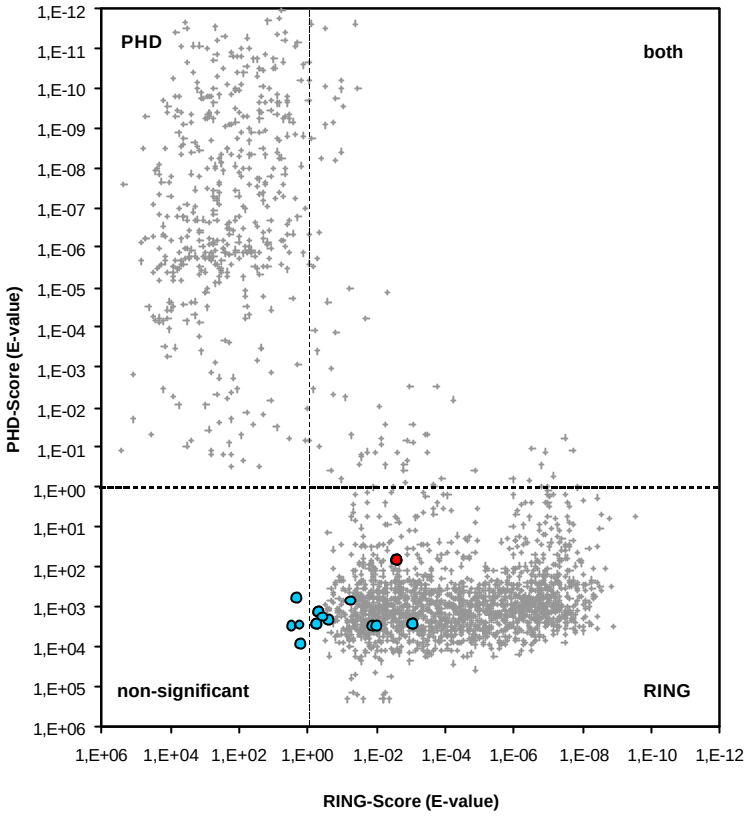


Figure 3-4 The score distribution of RING finger and PHD finger profiles is shown. Each protein region having significant or near-significant similarity to RING finger or PHD finger domains was scored against both discrimination profiles mentioned in the text. The axes indicate the E-values as a measure of significance. A dashed line indicates an E-value of 1, tentatively separating positive and negative predictions. Classic RING finger and PHD domains are evident as two separated clusters, with only a few sequences having significant scores with both profiles. The group of viral proteins discussed in 3.4.4 is indicated by blue circles, and the zinc finger domain of MEKK1 as a red circle. Both protein classes are members of the RING finger cluster. Some of the viral proteins do not reach significant RING scores with the given profile (blue circles left of the dashed line). A more sensitive profile, which already includes the significant viral proteins in the training set, would yield convincing RING scores for all of the classification targets.

3.4.1.2 Role of zinc-coordinating residues for subfamily determination

The classic RING finger is a zinc-finger structure coordinating two zinc cations in a cross-braced arrangement. For zinc-coordinating, eight cysteine and histidine residues are either ordered as C3H2C3 (RING-H2) or as C3HC4 (RING-HC). Besides this arrangement of the coordinating residues, there are several proteins with RING-finger like domains displaying a different arrangement of these residues or lack even some of them. However, sequence comparisons suggest an evolutionary relationship for these proteins to classic RING finger proteins. Like the classic RING fingers, several RING-finger like proteins from different families have been shown to function as ligases for Ub (TRIAD/RBR/Parkin triad family and U-box family) and in some cases for SUMO (PIAS family) (Hatakeyama, 2001, Johnson, 2001, Marin, 2004). Therefore, a grouping of classic RING-finger proteins and RING-finger like

proteins seems reasonable. To find the whole protein set of each subfamily, subfamily-specific profiles had to be generated.

The alignment in Figure 3-5 shows the RING finger domain of representative members of the classic RING finger family and all RING finger variants. Obviously, the classic RING finger domain is defined by a series of eight well conserved histidines and cysteines, which coordinate the zinc ions and already constitute large parts of the domain. The members of the U-box family deviate most strikingly from the classic RING finger, as the U-box lacks all zinc-coordinating residues (Aravind et al., 2000). Other families have lost only few of these residues or use different residues for zinc-binding at certain positions like the PIAS family or even more pronounced in some proteins termed 'degenerated RING fingers' here.

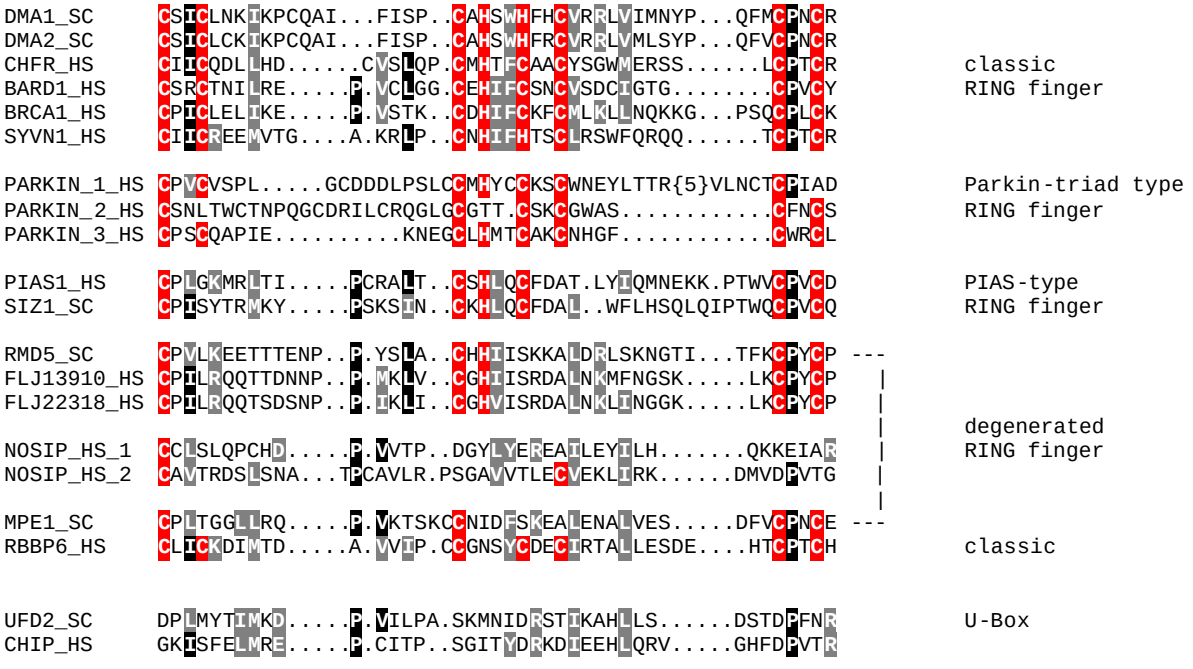


Figure 3-5 Alignment of representative members of each RING finger subfamily including the zinc less U-box domain. The classic RING finger proteins are shown at the top, while RING finger variants are shown below. Red background indicates zinc coordinating residues or potentially zinc binding residues at positions homologous to those in the classic RING finger.

3.4.1.3 Overview of the RING superfamily including U-Box proteins

Altogether, 48 proteins in yeast and ~300 proteins in human were detected with the classic RING finger profile or one of its variants including the U-box (see Table 3-8). The classic RING finger accounts for 39 proteins in yeast and ~270 proteins in human and constitutes the largest group by far. Due to the size of the classic RING finger family, a complete list of its members is provided in the appendix. The family next in size is the Parkin finger triad family. Here, two yeast representatives are opposed to 14 in human.

Not all members of the RING superfamily are Ub-specific E3s, but some ligate other type I Ub-like modifiers to substrates. The PIAS family comprises the SUMO-ligases Siz1 and Siz2 in yeast (Johnson et al., 2001, Takahashi, 2003), as well as Mms21 that has been assigned a Ub-ligase activity (Hofmann et al., 1999). Interestingly, the *S. pombe* orthologue of Mms21, Nse2/Pli2 (P87298) has recently been reported to be a SUMO ligase (Andrews, 2005). In human, seven proteins with the PIAS-type RING finger are expressed, with PIAS1, PIAS2/PIASx, PIAS3 and PIAS4/PIASy being known SUMO-ligases (Dohmen, 2004).

For the U-box family, Prp19 and Ufd2 were found in yeast. Both have been reported to have Ub-ligase activity (Koegl et al., 1999, Ohi, 2003). Of the six U-box homologues in human, CHIP is the best studied one with a well explored Ub-ligase activity depending on chaperones as substrate specificity factors (Connell et al., 2001).

Table 3-8 Overview of the RING family including UBox proteins, the HECT family and the A20 zinc finger family.

	Yeast	with orthologues	Human
classic RING	39	27	264
PIAS type RING	2	3	14
Parkin finger triad	3	2	7
degenerate RING	2	2	6
U-Box	2	2	5
HECT	5	5	28
A20 zinc finger	-	-	7

3.4.2 Comparing the RING finger and its variants

3.4.2.1 The Parkin finger triad

According to sequence analysis done here, the Parkin triad family consists of three consecutive copies of complex zinc fingers ('Parkin finger') with similarity to the classic RING finger rather than of two RING finger copies with a distinct zinc finger in between (Marin et al., 2004). On closer examination of the first two zinc finger domains within the Parkin finger triad of Parkin, individual residues that normally bind zinc have been replaced by non-binding ones. By contrast, the third RING finger still holds all eight zinc coordinating residues and, like the first RING finger domain, is detected with significant scores by the classic RING finger profile. As the intermediate RING finger copy lacks two zinc coordinating residues and obviously requires a gap to align the remaining cysteines correctly as seen in Figure 3-5, it is not detected significantly by the classic RING finger profile. Nevertheless, when regarding conserved residues outside the zinc-coordinating cysteine and histidine residues, the relationship between the intermediated zinc finger and the RING finger becomes clearer.

Table 3-9 Yeast and human Parkin finger triad proteins. A subtable contains orthology assignments. Gene names in italics denote pseudogenes.

Parkin finger triad					
Yeast		Human			
Gene name	ORF	Gene name	Uniprot number	Gene name	Uniprot number
YKR017C	YKR017C	ANKIB1	Q9P2G1	PARK2	O60260
ITT1	YML068W	ARIH1/UBCH7BP	Q9Y4X5	RNF14/ARA54	Q9UBS8
Assignment of orthologues		ARIH2/TRIAD1	O95376	RNF144	P50876
Yeast	**Human**	IBRDC1	Q8TC41	RNF19	Q9NV58
YKR017C	ARIH1/UBCH7BP	IBRDC2	Q7Z419	RNF31	Q96EP0
ITT1	RNF14/ARA54	IBRDC3	Q6ZMZ0	UBCE7IP1/TRIAD3	Q9NWF9
		PARC	Q8IWT3	UBCE7IP3/C20ORF18	Q9BYM8
				GENSCAN00000039330H	

3.4.2.2 PIAS-type RING finger

In PIAS1 and Siz1, only six of the zinc coordinating residues are left compared to the classic RING finger. PIAS3 lacks even three of these residues. The zinc finger domains in PIAS family members are detectable using the classic RING finger profile, but with weak scores below threshold. This observation is probably due to the high weight of zinc coordinating residues during profile construction. As described above, a PIAS specific profile of the RING finger domain had to be generated to find all PIAS family members (Table 3-10).

Table 3-10 Yeast and human PIAS-type RING finger proteins. A subtable contains orthology assignments.

RF_PIAS			
Yeast		Human	
Gene name	ORF	Gene name	Uniprot number
SIZ1	YDR409W	FLJ32440	Q96MF7
NFI1/SIZ2	YOR156C	PIAS1	O75925
MMS21	YEL019C	PIAS2/PIASx	O75928
Assignment of orthologues		PIAS3	Q9Y6X2
Yeast	**Human**	PIAS4/PIASy	Q8N2W9
SIZ1, NFI1/SIZ2	PIAS1-4	RAI17	Q9ULJ6
MMS21	FLJ32440	ZIMP7	Q8NF64

3.4.2.3 Highly degenerated RING-finger

Within this work, several proteins were found that exhibit a weak similarity to the classic RING finger. This similarity has so far not been described in the literature. However, control profiles could confirm a homologous relationship between the classic RING finger and these proteins. The degeneration of the series of zinc binding residues in these proteins is even more pronounced than in the protein families described so far and they are abstractly denoted as 'degenerated RING fingers' here (see Table 3-11). A general role of these atypical RING fingers as E3 has still to be established, although at least Rmd5 is required for the ubiquitylation of the gluconeogenetic enzyme fructose-1,6-bisphosphatase (Regelmann, 2003). Rmd5 and its human orthologues FLJ13910 and FLJ22318 share a set of rudimentary cysteine and histidine residues used for zinc coordination in classic RING finger proteins. Interestingly, the

degeneration of the zinc- binding residues in yeast Mpe1 and its putative human orthologue RBBP6 seems to have progressed at different rates. While RBBP6 may be classified as classic RING finger protein, Mpe1 is lacking three zinc-coordinating residues. It should be mentioned that RBBP6 is markedly larger than Mpe1 and likely has new or additional functions. In this respect, the different conservation of the RING fingers may indicate a change in function of the RING finger and the corresponding protein. In the case of human NOSIP with two degenerated RING finger domains, no obvious yeast orthologue could be assigned, although Slx8 may be a candidate. By contrast, in *N. crassa* and *F. gramineum* NOSIP orthologues seem to exist. In those proteins, the RING finger domains are more closely related to the classic RING finger than the human NOSIP RING finger domains.

Table 3-11 Yeast and human degenerated RING finger proteins. A subtable contains orthology assignments.

degenerated RING finger			
Yeast		**Human**	
Gene name	**ORF**	**Gene name**	**Uniprot number**
MPE1	YKL059C	C20orf43	Q9BY42
FYV10	YIL097W	FLJ13910	Q9H871
RMD5	YDR255C	FLJ22318	Q96G75
Assignment of orthologues		MAEA	Q9BQ11
Yeast	**Human**	NOSIP	Q96FD2
MPE1	RBBP6	PPIL2	Q13356
RMD5	FLJ22318, FLJ13910		
FYV10	MAEA		

3.4.2.4 No zinc binding capabilities, but still E3 activity: U-box

The U-box family is an extreme example for the degeneration of zinc-coordinating residues of the RING finger domain. For example, in human CHIP and yeast Ufd2, none of these residues is retained (see Figure 3-5). However, the U-box is still able to adopt the same structure as several structurally solved RING domains (Ohi et al., 2003). Both the U-box and the RING finger domain share a central α-helix surrounded by short β-strands and these secondary structure elements can almost be completely superimposed. The U-box fold is mainly stabilized by a hydrophobic core with residues occupying the originally zinc chelating positions, which have a strong impact for the correct fold (Ohi et al., 2003). Besides structural similarity, other features shared between U-Box proteins and classic RING finger proteins are the ability to bind to E2s and to act as Ub ligases (Pringa et al., 2001) (Jiang et al., 2001).

Table 3-12 Yeast and human U-box proteins. A subtable contains orthology assignments. A gene name in italics denotes a pseudogene.

U-Box Yeast Gene name	ORF	Human Gene name	Uniprot number
UFD2	YDL190C	CHIP/STUB1	Q9UNE7
PRP19	YLL036C	PRP19/SNEV	Q9UMS4
Assignment of orthologues		UBE4A	Q14139
Yeast	**Human**	UBE4B/UFD2	O95155
UFD2	UBE4B/UFD2	WDSAM1	Q8N6N8
PRP19	PRP19/SNEV	*GENSCAN00000045262H*	

3.4.3 Assignments of orthologues for the RING-type ligases

41 out of the 53 yeast RING superfamily members could be assigned to one or more human orthologues and only in a few cases several human paralogues had to be assigned to one yeast protein (see Table 3-13). Some special cases will be described here in more detail.

Table 3-13 Assignment of yeast/human orthologues for classic RING finger. For orthology assignments of non-classic RING finger families see previous tables. See appendix for ORFs or Uniprot numbers. Commas separate different proteins, backslashes alternative names.

Assignment of orthologues Yeast	Human	Yeast	Human
APC11	ANAPC11	PEX10	PEX10
BRE1	RNF20, RNF40/KIAA0661	PEX12	PEX12
CWC24	ZNF183, ZNF183L1	RAD18	RAD18
DMA1, DMA2	CHFR, RNF8	RAD5, RAD16, RIS1	SMARCA3/HIP116
FAP1	NFX1	SSM4	MARCH6/KIAA0597
HRD1	SYVN1/HRD1	TFB3	MNAT1
HRT1	RBX1	VPS8	KIAA0804
MAG2	RIE2	YDR128W	FLJ12270/KIAA1923
MOT2	CNOT4	YDR266C	ZNF598
PEP3	VPS18	YHL010C	BRAP
PEP5	VPS11	YMR247C	ZNF294
		UBR1, UBR2	UBR1, UBR2, UBR1L1/FLJ45053

3.4.3.1 Dma1/Dma2 vs. CHFR/RNF8

Difficulties were encountered finding orthologues for yeast Dma1/Dma2. Two alternative human proteins (CHFR and RNF18) appeared to be suitable candidates. The closely related Dma1 and Dma2 are classic RING-H2 finger proteins and involved in regulating spindle position and orientation. Besides the RING finger domain, there is a FHA domain (Forkhead-associated) present in all four proteins upstream of the RING finger (see Figure 3-6). The FHA domain is a phosphopeptide-binding motif often found in regulatory proteins (Hofmann, 1995) (Durocher, 1999). All four proteins are the only examples for a combined FHA/RING domain topology in yeast and human. CHFR corresponds to Dma1/Dma2 functionally (Fraschini, 2004).

However, sequence comparison favours RNF8 as Dma1/Dma2 orthologue when analyzing the FHA domain only. This can be seen from 43% sequence similarity between RNF8 and each Dma1 and Dma2 versus ~30% between CHFR and Dma1/Dma2. At the same time, the RING finger domains of both human sequences are more similar to each other than to the homologous domain in Dma1/Dma2, i.e. the human genes seem to have arisen from gene duplication after the splitting of the metazoan and fungal lineage. A 2:2 relationship is the most probable explanation for the evolutionary history of this family, i.e. Dma1/Dma2 are assumed to be co-orthologous to CHFR/RNF8 (see Table 3-13).

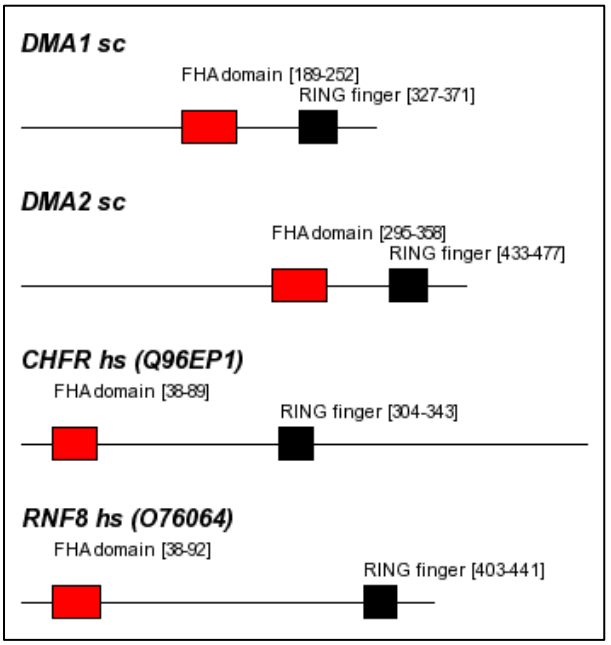


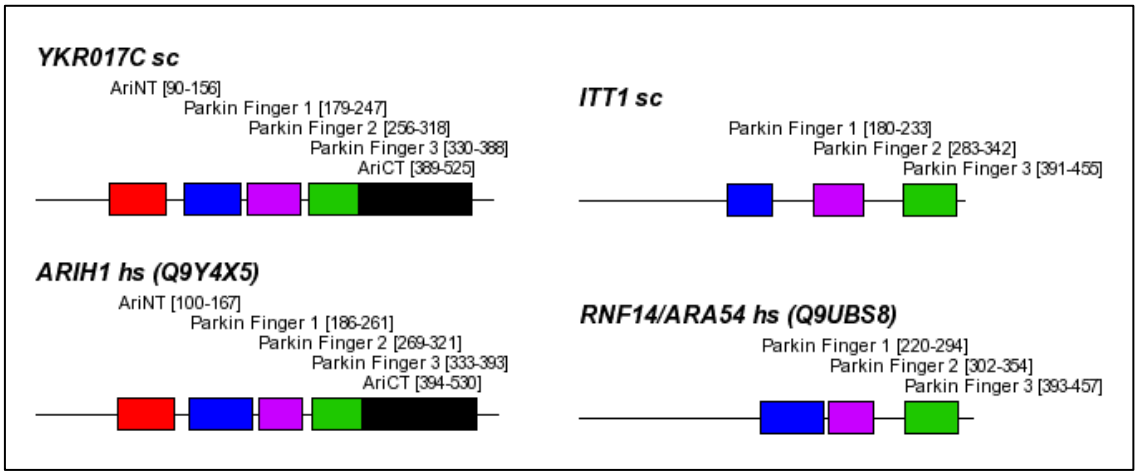
Figure 3-6 Domain structure of Dma1, Dma1, CHFR and RNF8.

3.4.3.2 YKR017C vs. ARIH1/UBCH7BP

YKR017C is a protein with a Parkin finger triad, a motif also found in Parkin, the causative gene of Parkinson disease (Tanaka, 2004). Interestingly, YKR017C enhanced the toxicity of ectopically expressed human huntingtin exon 1 with an expanded poly(Q) repeat in yeast (Willingham, 2003). Both Parkinson disease and Huntington disease are characterized by ubiquitylated intraneuronal inclusion bodies in affected brain cells, but differ in the cellular mechanisms leading to pathogenesis (Willingham et al., 2003). Therefore, the Parkin triad motif links at least two proteins with effect on toxicity of poly(Q)-expanded proteins. In the publication of Willingham et al., YKR017C has been mentioned to be involved in Ub-mediated protein degradation, although no further reference is given for this claim (Willingham et al., 2003).

Parkin is a known Ub ligase with its Parkin finger triad being crucial for its function. Mutations in the Parkin finger triad of the Parkin gene have been detected in Parkinson disease patients (Tanaka et al., 2004) suggesting that the homologous region in YKR017C may play a role for the observations of Willingham et al.. However, YKR017C exhibits an additional N-terminal domain ('Ariadne N-terminal domain'), which is related to the UBA-domain and therefore probably has a Ub-binding capacity that might be implicated in the observations made for YKR017C as well (see Figure 3-7). The human orthologue of YKR017C is ARIH1 that has been shown to bind to the E2s UBCH7 and UBCH8 (Moynihan, 1999). Thus, there are several hints that functionally link YKR017C to the UPS, most likely as a Ub ligase. Notably, no indications for a role of ARIH1 in Parkinson or Huntington disease have been reported so far. Instead, ARIH1 appears to be involved in Ub-dependent degradation of a protein involved in protein translation, 4EHP (Tan, 2003).

Figure 3-7 Domain structure of yeast YKR017C and ITT1 as well as of their human orthologues. The individual



Parkin fingers are shown, termed 'Parkin Finger 1-3'.

3.4.4 No evidence for PHD fingers as Ubiquitin ligases

Note: the following work has been published as *"No evidence for PHD fingers as Ubiquitin ligases."* by Scheel, H. and K. Hofmann in Trends Cell Biol 13(6): 285-7 (Scheel, 2003).

In a 2003 article, Coscoy and Ganem had proposed a third class of Ub ligases, which rely on PHD fingers instead of RING finger domains (Coscoy, 2003). As mentioned above, PHD fingers are complex zinc fingers like RING fingers and typically occur in proteins involved in chromatin regulation. The proposal of Coscoy and Ganem was based on two different protein families that reportedly contain PHD fingers and have been shown to possess Ub ligase activity: the mitogen-activated protein kinase kinase kinase MEKK1 (Lu, 2002) and a group of viral proteins including MIR1 and MIR2 encoded by the Kaposi sarcoma virus KSHV (Boname, 2001, Coscoy, 2001, Fruh, 2002, Mansouri, 2003). While there is little doubt that

both protein classes can catalyse the transfer of Ub, the notion that the zinc fingers in those proteins should be classified as PHD fingers must be disagreed on. According to sequence analysis in this work, both protein classes are slightly atypical members of the RING finger family. To show this, the sequences were compared against both the pure RING finger and PHD finger profiles from 3.4.1. Both MEKK1 and members of the viral MIR family yielded better scores with the RING finger profile and unambiguously grouped with the RING fingers in Figure 3-4. Similar results were obtained by control profiles constructed from the MIR family, which significantly retrieved several RING finger proteins but no PHD protein. Thus, in 2003 there was no reason to assume that true PHD finger proteins are involved in ubiquitylation. Similar conclusions have been drawn by Aravind et al. (Aravind, 2003).

In a more recent publication, a true PHD finger protein, AIRE (O43918), has been shown to have E3 activity (Uchida, 2004). PHD finger proteins should therefore not be ruled out as E3 ligases, but AIRE remains the only PHD protein with this activity so far. The putative PHD finger proteins with E3 activity analysed in this chapter clearly belong to the RING finger family and therefore do not contribute to a speculative new class of PHD-type E3 ligases.

3.4.5 RING-cullin based E3s

3.4.5.1 Few scaffolds for RING-cullin based E3s

As described in the introduction, complex E3s are composed of modules, which may roughly be subdivided into a RING-cullin based scaffold and different adaptors mediating substrate specificity (see Figure 1-3 and Table 3-14). The number of distinct RING and cullin proteins in complex E3s is manageable with just 2 or 3 different RING finger proteins and 3 to 8 different cullins in yeast or man, respectively (see Table 3-15).

Table 3-14 RING-cullin based E3s.

human complex	cullin	RING	adaptor	substrate binding
SCF	CUL1	RBX1	SKP1	F-Box protein
VBC, ECS, SCF2	CUL2	RBX1	TCEB1/Elongin C TCEB2/Elongin B	SOCS-Box protein
BCR3, SCF3	CUL3	RBX1	BTB	BTB
VDC, SCF4	CUL4A	RBX1	DDB1	
?	CUL4B	RBX1	?	?
SCF5	CUL5	RNF7/ROC2	TCEB1/Elongin C TCEB2/Elongin B	SOCS-Box protein
SCF7	CUL7	RBX1	SKP1	F-Box protein
APC	ANAPC2	ANAPC11	multiple	
?	PARC	?	?	?
yeast complex	**cullin**	**RING protein**	**adaptor**	**substrate binding**
SCF	CDC53	HRT1	SKP1	F-Box protein
?	?	HRT1	ELC1	ELA1, RAD7
APC	APC2	APC11	mulitple	CDC20, HCT1
?	CUL3	HRT1(?)	BTB	BTB

Table 3-15 List of cullins and SKP1-like proteins in yeast and human. Pseudogenes are printed in italics.

Cullin			
Yeast		**Human**	
Gene name	**ORF**	**Gene name**	**Uniprot number**
APC2	YLR127C	ANAPC2	Q9UJX6
CDC53	YDL132W	CUL1	Q13616
CUL3	YGR003W	CUL2	Q13617
RTT101	YJL047C	CUL3	Q13618
Assignment of orthologues		CUL4A	Q13619
Yeast	**Human**	CUL4B	Q13620
APC2	ANAPC2	CUL5	Q93034
CDC53	CUL1, CUL2	CUL7	Q14999
CUL3	CUL3, CUL4A, CUL4B	PARC	Q8IWT3
SKP1-like			
Yeast		**Human**	
Gene name	**ORF**	**Gene name**	**Uniprot number**
SKP1	YDR328C	SKP1	P63208
ELC1	YPL046C	TCEB1/Elongin C	Q15369
		-	*P78561*
		-	*P78389*
		RP1-254P11.1-001	*Q9H575*
		Fos39347_1	*O75863*

3.4.5.2 Defining the substrate-binding subunits

RING-cullin based complex E3s use a variety of substrate-binding subunits. For example, in the SCF complex the substrate specificity is conferred by an F-Box protein. The VBC complex consists mainly of the RING-cullin scaffold and Elongin C, which in turn binds to proteins with a SOCS domain, e.g. VHL. The SOCS domain proteins have a function

analogous to the F-Box subunit in the SCF complex. The third major class of RING-cullin based E3s employs BTB domain proteins as substrate-binding subunits. The BTB proteins are somewhat different from F-Box and SOCS proteins, as they directly bind to the RING-cullin scaffold while F-Box and SOCS proteins are connected to the scaffold via the adaptors Skp1 and Elongin C, respectively. Interestingly, the structure of the BTB domain has a fold that is similar to Skp1 and Elongin C (Ahmad, 1998). In order to compare the substrate-binding subunits of yeast and human, F-Box, SOCS and BTB protein sets were determined by profile-based methods.

F-Box, SOCS and BTB proteins are present in both yeast and human. Each of these three classes was subclassified on the basis of the distinct protein-protein-interaction domains that are present in each protein. Again, profiles were used for this secondary classification step, including profiles based on the WD40 repeat, the KELCH domain, leucine-rich repeats (LRR), the Ankyrin motif, zinc finger motifs, etc.. A comprehensive list of the substrate-binding subunits is provided in the appendix.

One specific complex E3 ligase is the APC, which is not considered here. A novel type of RING-cullin based E3s employs DDB1 as a substrate-binding subunit and Cul4A/RBX1 as a scaffold (Wertz et al., 2004). This E3 is a rather specific example for a RING-cullin based E3 and a more general role for DDB1 based adaptors is yet unexplored. A yeast orthologue for human DDB1 could not be found.

3.4.5.3 Large families of potential substrate binding subunits

Substrate-binding proteins form large families both in yeast (28) and in human (237) (see Table 3-16). Within yeast, the F-Box family (20) contributes clearly more substrate-binding subunits than the SOCS (2) or BTB family (6). By contrast, in human each substrate binding class is significantly expanded with the BTB class accounting for 126 members, while the F-Box class and SOCS class display half that size with 67 and 41 members, respectively. The novel DDB1 class has just three members.

Not all proteins harbouring the typical substrate-binding subunits are guaranteed to be part of active complex E3s. For example, the yeast CBF3-kinetochore associated F-Box-protein Ctf13 interacts with Skp1 without any cullin-based E3 activity involved for proper function (Connelly, 1996). Moreover, the dimer of Skp1 and Rcy1 has a role in vesicular trafficking without being part of an active E3 ligase (Galan, 2001). Comparable examples are likely to occur in the BTB and SOCS classes as well.

Table 3-16 Substrate binding subunits.

adaptor type	substrate binding site	yeast	human
F-Box	**all**	**20**	**69**
	LRR	4	23
	WD40	2	10
	ZINC_FINGER	0	3
	KELCH	0	1
	GRF_RCC	1	1
	others/none	13	31
SOCS-Box	**all**	**2**	**41**
	ANKYRIN	0	17
	SH2	0	8
	SPRY	0	4
	WD40	0	3
	LRR	1	1
	others/none	1	8
BTB	**all**	**6**	**126**
	KELCH	2	48
	ZINC_FINGER	0	48
	ANKYRIN	1	5
	GRF_RCC	0	3
	MATH	0	2
	others/none	3	21

3.4.5.4 Assignment of orthologues

The adaptor proteins SKP1 and Elongin C have clear orthologues in both species and most yeast cullin proteins could be assigned to human orthologues except for RTT101 (see table Table 3-15). The assignment of orthologues for the F-Box, SOCS-Box and BTB proteins proved to be more difficult and only few clearly orthologous pairs were found (see Table 3-17). For example, the well known yeast F-Box Cdc4, which is required for G1/S and G2/M transition (Goh, 1999) has a true human orthologue (FBXW7) with similar functions (Strohmaier, 2001). Moreover, the human genome encodes three paralogous Elongin A proteins, which all are orthologous to the yeast SOCS-Box protein Ela1 (Elongin A) involved in transcription elongation (Botuyan, 1999). All three human proteins seem to be functionally similar and can form heterotrimers with Elongin B and C (Yamazaki, 2002).

Table 3-17 Orthology assignments for substrate-binding subunits of complex E3s.

Yeast	Human	Comments	Molecular Function / Biological process
CDC4	FBXW7	F-Box	cell cycle
HRT3	FBXO9	F-Box	nuclear Ub ligase
ELA1	TCEB3, TCEB3B, TCEB3C	SOCS-Box	RNA elongation
YIL001W	ABTB1	BTB	translation elongation
RAD7	LOC196394	SOCS-Box	nucleotide-excision repair, DNA damage recognition

3.4.5.5 Five novel BTB proteins in yeast

While screening the yeast database with a BTB-based profile, only Yil001w was matched significantly as compared to 126 in human. Besides this hit, several slightly subsignificant matches were also found and therefore considered as candidates for novel BTB

proteins. Only proteins with a sufficiently long match were selected for a subsequent validation step. Control profiles were then constructed from these candidates and their clear orthologues from other fungal organisms. As a result, five of the candidates turned out to be distant BTB homologues (see Figure 3-8). While the classic BTB protein Yil001w has a human orthologue, ABTB1/BPOZ, all five novel BTB proteins are specific to fungi, sometimes even without a clear orthologue in *S. pombe*. The individual BTB proteins will be discussed below.

K-channel tetramerization domain vs. BTB domain

The family of BTB proteins consists of two closely related subfamilies. Within the Pfam database, these subfamilies are described by two different classes of profile HMMs, called 'BTB' and 'K-channel tetramerization domain'. Comparing the structures that are available for each subfamily demonstrates a structural relationship. Therefore, good evidence exist that both subfamilies and the corresponding domains, respectively, are evolutionarily related. The name 'K-channel tetramerization domain' is somewhat misleading, because substrate-binding subunits of a complex E3 and channel proteins obviously share distinct functions. The 'K-channel tetramerization domain' has a functional motif inserted, which supports tetramerization, while the core domain with a BTB-like fold is not involved in the latter.

Ylr108c/Ydr132c

The previous short introduction to the BTB superfamily is important for understanding the analysis of Ylr108c and its close relative Ydr132c, two of the five new yeast ORFs with BTB-like domains (see Figure 3-8). A reverse profile search starting from the putative BTB-like regions in these proteins found members of the 'K-channel tetramerization domain' family. Clear orthologues were identified in other fungal species, *C. albicans* and *N. crassa*, while the best match in metazoan sequence databases was KCNV1 from mouse, a K-channel. However, the residues needed for tetramerization are absent in both proteins, so there is no reason to assume a function as K-channels. According to the profile-based searches, each ORF harbours two copies of the BTB domain, which is in good agreement with the classic BTB protein Yil001w. Interestingly, Ylr108c seems to interact with yeast Cul3, thus the BTB-like domains might be functional (Pintard, 2003). Both Ylr108c and Ydr132c lack a characteristic protein-protein interaction domain as it is found in other BTB proteins, e.g. KELCH or Ankyrin repeats, so a role as substrate binding subunits in a Cul3-based complex E3 remains elusive.

Mds3/Pmd1

Mds3 and Pmd1 (paralogue of Mds3) are more closely related to the proper BTB family and both contain KELCH-domains upstream of a singular BTB domain. Nonetheless, the similarity to the classic BTB domain is weak and only in the context of the KELCH-domains,

which are quite common among real BTB proteins, these proteins should be ranked as BTB proteins. The high divergence of both proteins might be the reason why even in *S. pombe* no orthologue could be defined. Interestingly, both novel BTB proteins are functionally linked, as they both are involved in the regulation of sporulation-specific genes (Benni, 1997).

Whi2

The fifth newly identified yeast BTB protein is Whi2, which like Ylr108c and Ydr132c belongs to the 'K-channel tetramerization domain' subfamily. Again, no known protein interaction domain was found in this protein. However, results from sequence analysis point to a valid, singular BTB domain in the case of Whi2. Whi2 has orthologues in *C. albicans*, *S. pombe* and *N. crassa*, but not in human.

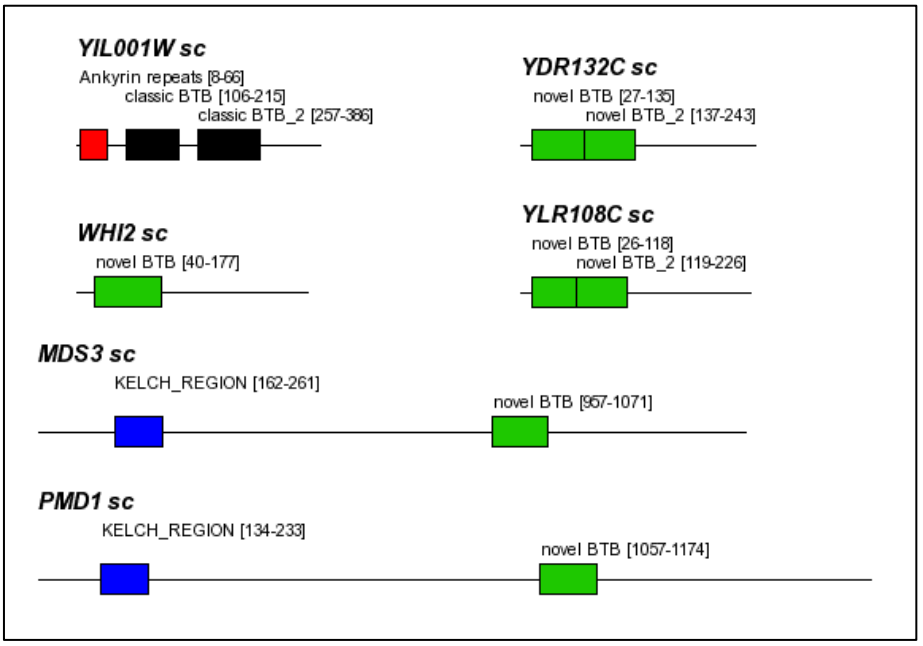


Figure 3-8 Domain structure of yeast BTB proteins.

3.4.6 HECT type Ub-ligases

HECT type ligases are shown in Table 3-18. The overall number of human HECT ligases is 28 and therefore much higher than in yeast, which encodes five HECT proteins. All five yeast HECT ligases could be assigned to human orthologues. In three cases, two or more human genes correspond to one yeast gene.

Table 3-18 HECT-type ligases.

HECT Yeast		Human			
Gene name	ORF	Gene name	Uniprot	Gene name	Uniprot
TOM1	YDR457W	EDD/HYD	O95071	HECW2/KIAA1301	Q9P2P5
RSP5	YER125W	HACE1/KIAA1320	Q5VU99	NEDD4	P46934
HUL5	YGL141W	HECTD1	Q9ULT8	NEDD4L	Q7Z5F1
HUL4	YJR036C	HECTD2	Q5U5R9	HECW1/NEDL1	Q9HCC7
UFD4	YKL010C	HERC2	O95714	HERC1/P532	Q15751
Assignment of orthologues		HERC3	Q15034	FLJ21156	Q5T447
Yeast	Human	HERC4	Q5VXS9	SMURF1	Q9HCE7
TOM1	UREB1	HERC5/CEBP1	Q9UII4	SMURF2	Q9HAU4
RSP5	NEDD4, NEDD4L	HERC6	Q8IVU3	TRIP12	Q14669
HUL5	UBE3B, UBE3C	ITCH	Q96J02	UBE3A/E6AP	Q05086
HUL4	HERC5, HERC3,	KIAA0317	O15033	UBE3B	Q9BXZ4
	UBE3A, HECTD2,	KIAA0614/FLJ30092	Q9Y4D8	UREB1	Q7Z6Z7
	HERC4, HERC6	KIAA1333/FLJ20333	Q9NXC0	WWP1	Q9H0M0
UFD4	TRIP12	UBE3C/KIAA0010	Q15386	WWP2	O00308

3.4.7 A20-zinc-finger-type Ub-ligases

A20-zinc-finger type ligases are a recently discovered class of Ub ligases (Wertz et al., 2004). The zinc finger of TNFAIP3/A20 has been demonstrated to modify RIP1 (receptor interacting protein) with lysine-48-linked Ub chains resulting in RIP1 degradation. RIP1 degradation in turn inactivates the NF-κB signalling pathway (Zhang, 2000). Remarkably, A20 has an additional function as deubiquitylating enzyme due to an OTU domain. Indeed, A20 is able to cleave both lysine-48- and lysine-63-linked Ub chains (Evans, 2004, Wertz et al., 2004). This dual functionality of A20 allows this protein to change the linkage type of ubiquitylated RIP1 by cleaving the lysine-63-linked chain and subsequently adding a lysine-48-linked chain. Proteins with an A20-zing finger could not be detected in yeast, in contrast to seven members in human (see Table 3-19).

Table 3-19 Human A20 Zn-finger type Ub-E3.

Human A20 Zn finger type Ub-E3	
Gene name	Uniprot number
ZA20D1/Cezanne1	Q6GQQ9
ZA20D2	O76080
C15orf16/Cezanne2	Q8TE49
AWP1	Q9GZY3
TEX27	Q9H8U3
RABGEF1	Q9UJ41
TNFAIP3/A20	P21580

3.4.8 Non-RING based SUMO-ligases

RanBP2 (Ran binding protein 2) has been shown to be a SUMO ligase associated with the nuclear pore complex and mediates sumoylation of Sp100 (Pichler, 2002). It contains two copies of a so-called RanBP-repeat and sumoylation activity seems to depend on these copies. RanBP2 is currently the only protein with RanBP-repeats that has been reported to function as SUMO ligase. No yeast RanBP2 orthologue could be detected.

Sumoylated human RanGAP1 is able to bind to RanBP2 and therefore to associate with the nuclear pore complex in order to mediate nuclear protein import (Matunis et al., 1996). In yeast, RanGAP1 lacks the sumoylation motif needed for conjugation of SUMO to RanGAP1 in human (data not shown). Taken together, yeast lacks the two prerequisites used to bring RanGAP1 to the nuclear pore complex in human. Nonetheless, SUMO conjugation seems to be involved in nuclear protein import in yeast suggesting a different mechanism than the one used in human (Stade, 2002).

Another SUMO ligase not present in yeast is the polycomb group protein Pc2 (O00257). Kagey et al. have observed a strongly stimulating effect for Pc2 on sumoylation of CtBP, a transcriptional corepressor (Kagey, 2003). The mechanism of Pc2 as a SUMO ligase relies on recruiting both the substrate and Ubc9, the SUMO E2.

3.5 Ub-hydrolases (DUB) and desumoylating enzymes

3.5.1 UCH family

This family of DUBs preferably cleaves Ub monomers from Ub precursors. Only one yeast protein, Yuh1, belongs to the UCH family (see Table 3-20). It is orthologous to both human UCHL1 and UCHL3, but Yuh1 has a slightly higher similarity to UCHL3 and shares some functional properties with it (Linghu, 2002, Wada, 1998). Thus, Yuh1 and UCHL3 are probably functional orthologues.

UCHL1, a close homologue of UCHL3, is highly expressed in the brain and constitutes ~1% of brain protein content. It is of particular interest, as it is mutated in some cases of familial Parkinson disease and has also been described to be a component of Lewy bodies (Leroy, 1998, Polymeropoulos, 1997).

Another UCH protein with a proposed role in human disease is BAP1 (BRCA1 associated protein 1), which has a C-terminal extension required for interaction with the tumour suppressor BRCA1 (breast cancer associated gene 1). This interaction enhances BRCA1 mediated growth inhibition, probably through deubiquitylation and therefore stabilization (Jensen, 1998).

Table 3-20 Yeast and human UCH proteins. Yuh1 and UCHL3 are probably orthologues.

Yeast		Human	
Gene name	ORF	Gene name	Uniprot number
-	-	UCHL1	P09936
YUH1	YJR099W	UCHL3	P15374
-	-	UCHL5/UCH37	Q9Y5K5
-	-	BAP1	Q92560

3.5.2 USP family

3.5.2.1 USP-type DUBs are highly diversified

By means of the profile constructed from the catalytic domain, all USP members could be readily identified in yeast and human (see Table 3-21). The USP family is by far the largest family of DUBs with 58 members in human and 18 in yeast. Only few members in yeast and human lack the catalytic cysteine and probably have no deubiquitylating activity (see Table 3-21). Although the best-conserved part in the USP proteins is the catalytic domain, some USP proteins share detectable sequence conservation in the N- and C-terminal regions. Often, additional known domains accompany the catalytic domain and similar domain topologies are found in different USPs (see Figure 3-9). Therefore, sequence features outside the catalytic

domain could be used for assignment of orthologues besides a dendrogram analysis (Table 3-22).

Table 3-21 List of yeast and human USP members. USPs with a complete catalytic triad are underlined.

USP/UBP Yeast Gene name	**ORF**		
UBP1	YDL122W	UBP10	YNL186W
UBP2	YOR124C	UBP11	YKR098C
UBP3	YER151C	UPB12	YJL197W
DOA4	YDR069C	UBP13	YBL067C
UBP5	YER144C	UBP14	YBR058C
UBP6	YFR010W	UBP15	YMR304W
UBP7	YIL156W	UBP16	YPL072W
UBP8	YMR223W	PAN2	YGL094C
UBP9	YER098W	SAD1	YFR005C
Human Gene name	**accession number**	**Gene name**	**accession number**
USP1	O94782	USP28	Q96RU2
USP2	O75604	USP29	Q9HBJ7
USP3	Q9Y6I4	USP30	Q96JX4
USP4	Q13107	USP31/KIAA1203	Q70CQ4
USP5/ISOT	P45974	USP32	Q8NFA0
USP6	P35125	USP33	Q8TEY7
USP7/HAUSP	Q93009	USP34	O60316
USP8	P40818	USP35	Q9P2H5
USP9X	Q93008	USP36	Q9P275
USP9Y	O00507	USP37	Q86T82
USP10	Q14694	USP38	Q8NB14
USP11	P51784	USP39	Q96RK9
USP12	O75317	USP40	Q9NVE5
USP13/ISOT3	Q92995	USP41	Q70BM7
USP14	P54578	USP42	Q9H9J4
USP15	Q9Y4E8	USP43	Q70EL4
USP16	Q9Y5T5	USP44	Q9H0E7
USP17	Q6QN14	USP45	Q70EL2
USP17L	Q7RTZ2	USP46	Q80V95
USP17L2/DUB3	Q6R6M4	USP47/FLJ14456	Q96K76
USP18	Q9UMW8	USP48	Q86UV5
USP19	O94966	USP49	Q70CQ1
USP20	Q9Y2K6	USP50	Q70EL3
USP21	Q9UK80	USP51	Q70EK9
USP22	Q9UPT9	USP52/PAN2/KIAA0710	Q8IVE1
USP24	Q9UPU5	USP53	Q70EK8
USP25	Q9UHP3	USP54	Q70EL1
USP26	Q9BXU7	CYLD	Q9NQC7
USP27X	GENSCAN00000038249	C13orf22	Q14109

Table 3-22 Orthologous relationships between yeast and human USPs.

USP/UBP Orthologues Yeast	Human	additional domains
UBP1	-	
UBP2	USP25,USP28	
UBP3	USP10	
DOA4, UBP5, UBP7	USP8	Rhodanese
UBP6	USP14	Ub-like
UBP8	USP22, USP27X, USP51	ZF_UBP
UBP9,UBP13	USP12,USP46	
UBP10	-	
UBP11	-	
UBP12	USP4, USP11, USP15	HAUSP2A
UBP14	USP5,USP13	ZF_UBP
UBP15	USP7	MATH, HAUSP2A, HAUSP2B
UBP16	-	
PAN2	USP52	nuclease domain
SAD1	USP39	ZF_UBP

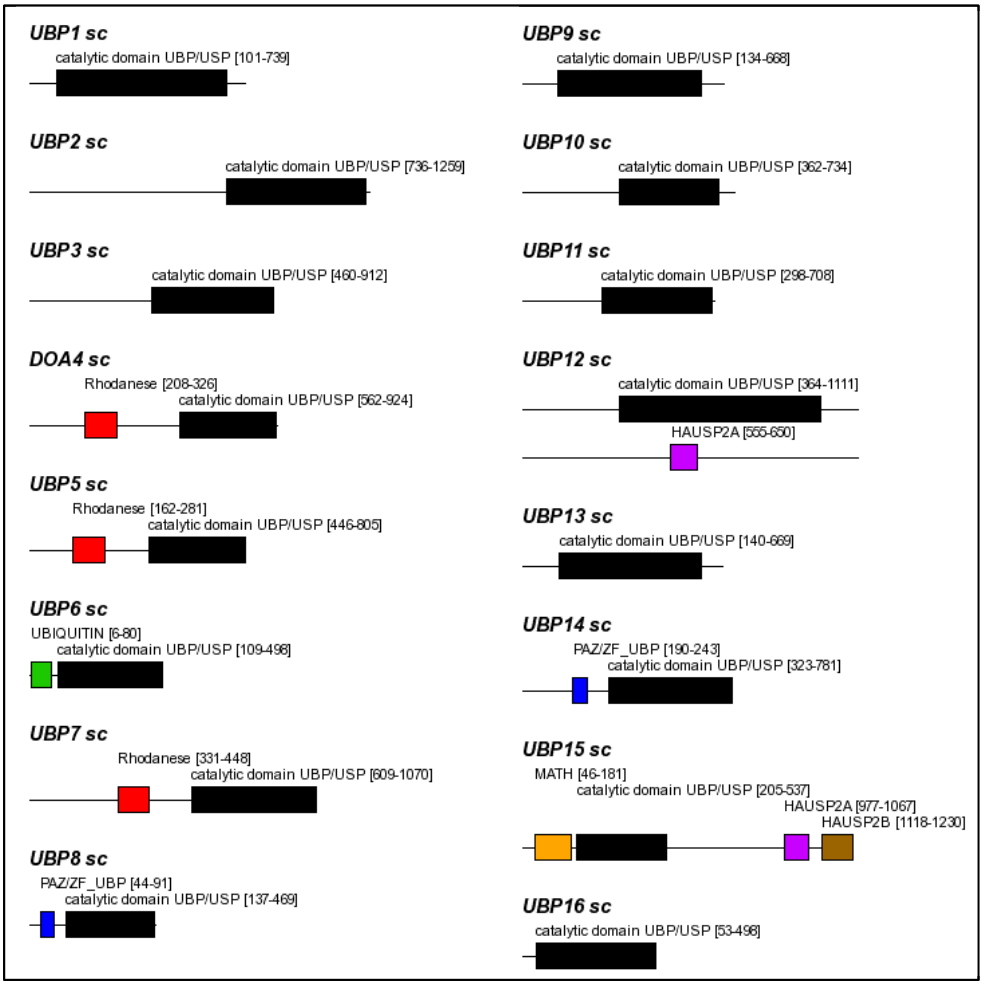
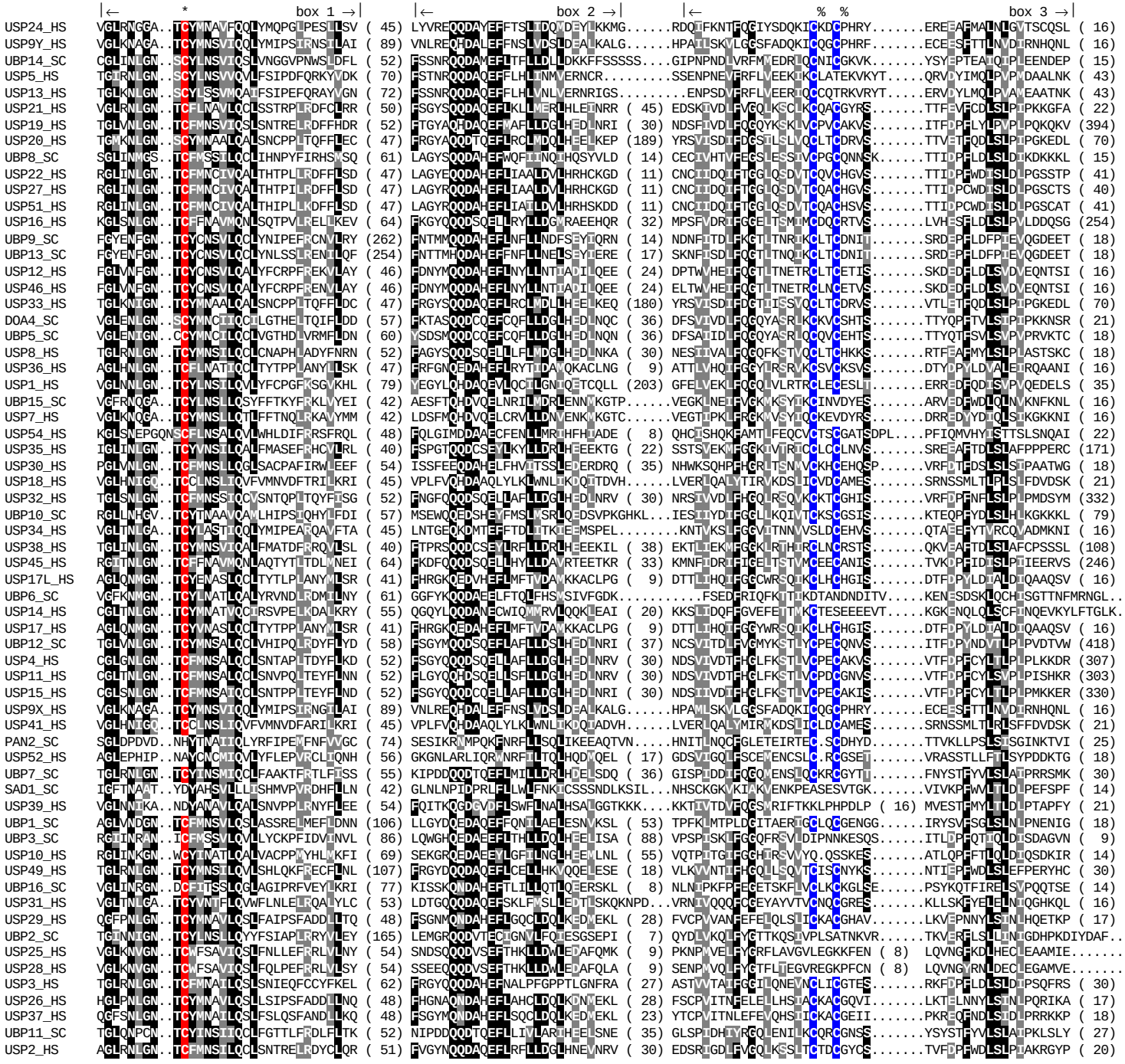


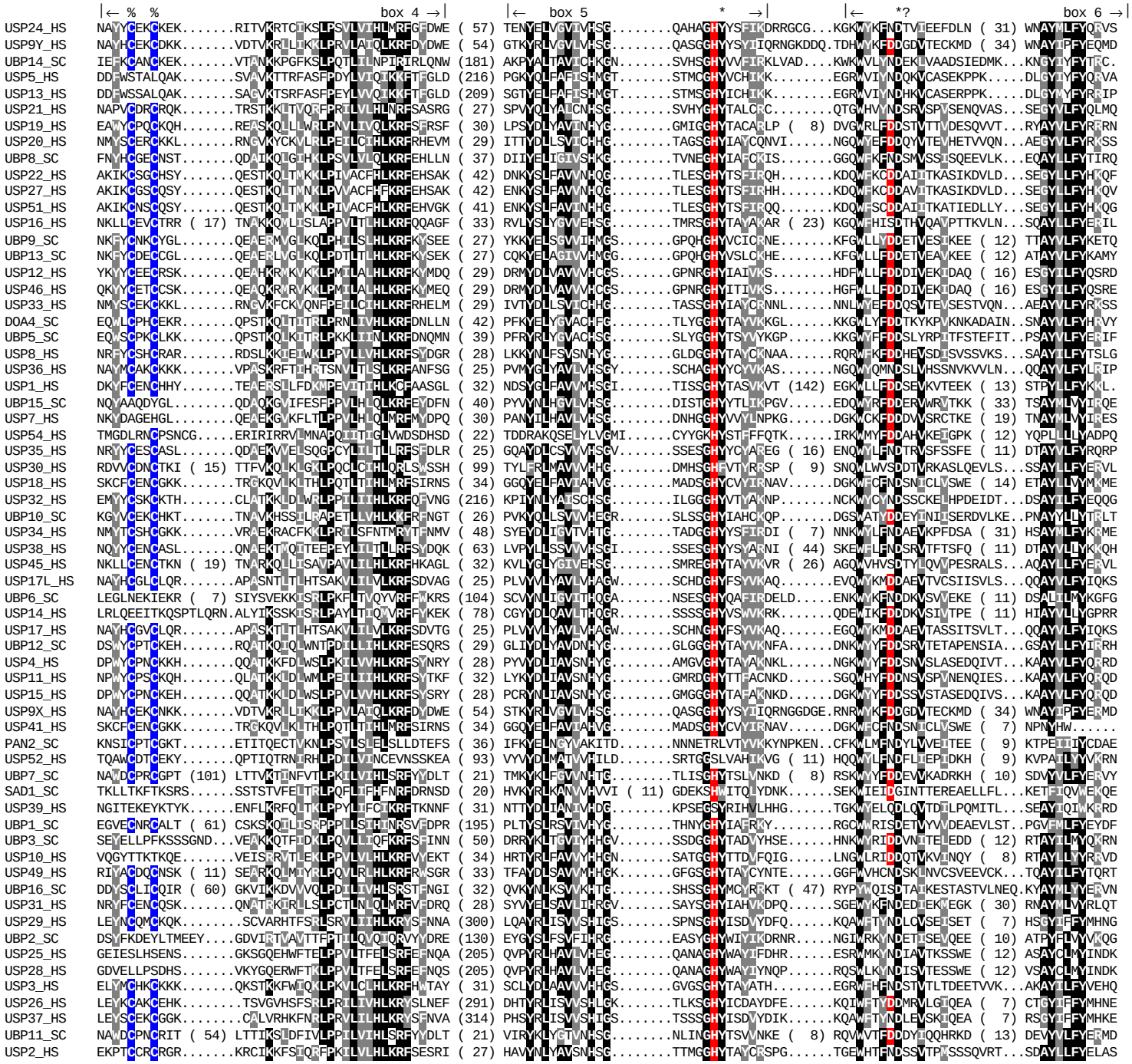
Figure 3-9 Domain structures of yeast USPs and the exact localization of their catalytic domain. The non-catalytic Sad1 and Pan2 proteins are not shown. Like its human orthologue USP15, yeast Ubp12 has a HAUSP2A domain inserted in the catalytic domain.

3.5.2.2 The catalytic domain consists of six major boxes

All USP-type DUBs share a rather large catalytic domain with an average length of ~350 residues. The boundaries of this domain as defined here are in good agreement with the X-ray structure of USP7/HAUSP, solved by Hu et al. (Hu, 2002). In this structure, two main functional elements can be recognized, an extensive Ub-binding surface and a catalytic cleft. The cleft is formed by two sequence boxes conserved in all USPs, a Cys-Box and a His-Box. From sequence analysis, the catalytic domain could be further refined yielding six major sequence boxes conserved across the USPs (see Figure 3-10). Block 1 contains the catalytic cysteine and corresponds to the Cys-Box. The His-Box was divided into two boxes that correspond to boxes 5 and 6. Here, box 5 carries the catalytic histidine, while box 6 harbours an invariant aspartate found in all catalytically active USPs. The residue N-terminally adjacent to this aspartate is frequently an asparagine or aspartate in other USPs. According to Hu et al., this preceding asparagine or aspartate is involved in catalysis (Hu et al., 2002). It is noteworthy that boxes 2, 3 and 4 are present in all USPs, although they do not contain any catalytical residue.

Figure 3-10 (next page) Alignment of yeast and human USPs. Conserved alignment positions are printed on black background. Positions occupied by amino acids with similar physicochemical properties are shaded in grey if supported by 50% of all available USP sequences. Residues of the catalytic triad are marked with an asterisk and printed on red background, while residues of the potential zinc finger are depicted on blue background and marked with a '%'. A '?' marks the invariant aspartate that could act as an alternative to the adjacent catalytic residue. Boxes are numbered from 1 to 6 and correspond to 'Box_1' to 'Box_6' in Figure 3-11.


|← * box 1 →| |← box 2 →| |← % % box 3 →|
USP24_HS VGLRNGCA..TCYMNAVFQQLYMQPGLPESLLSV (45) LYVREQQDAYEFFTSLIDQMDEYLKKMG.....RDQIFKNTFQGIYSDQKICKDCPHRY.....ERESAFMALNLGVTSCQSL (16)
USP9Y_HS VGLKNAGA..TCYMNSVIQQLYMIPSIRNSTLAAI (89) VNLREQHDALEFFNSLVDSIDEALKALG.....HPAIISKVLGGSFADQKICQGCPHRF.....ECESSFTTLNVDIRNHQNL (16)
UBP14_SC CGLINLGN..SCYLNSVIQSLVNGGVPNWSLDFL (52) FSSNRQQDAMEFLTFLLDLLDKKFFSSSSS....GIPNPNDLVRFMVEDRIQCNICGKVK.....YSYSPTEAIQIPLEENDEP (15)
USP5_HS TGIRNLGN..SCYLNSVVQVLFSIPDFQRKYDK (70) FSTNRQQDAQEFFLHLINMVERNCR.....SSENPNEVFRFLVEEKIKCLATEKVKYT.....QRVDYIMQLPVPVDAALNK (43)
USP13_HS TGLKNLGN..SCYLSSVMQALFSIPEFQRAYVGN (72) FSSNRQQDAQEFFLHLVNLVERNRIGS.....ENPSDVFRFLVEERIQCCQTRKVRYT.....ERVDYLMQLPVAVEAATNK (43)
USP21_HS VGLRNLGN..TCFLNAVLQCLSSTRPLFDFCLRR (50) FSGYSQQDAQEFLKLLVERLHLEINRR (45) EDSKIVDLFGVEQLKSCLKCQAQGYRS.....TTFIVFCDLSLPIPKKGFA (22)
USP19_HS TGLVNLGN..TCFMNSVIQSLSNTRELRDFFHDR (52) FTGYAQHDAQEFMAFLLDGLHEDLNRI (30) NDSFIVDLFGQGQYKSKLVCPVCAKVS.....ITFIPFLYLPVPLPQKQKV (394)
USP20_HS TGVKNLGN..SCYMNAALQALSNCPPLTQFFLEC (47) FRGYAQQDTQEFLRCLMDQLHEELKEP (189) YRSVISDIFDGSLLSLVQCLTCDRVS.....TTVETFQDLSLPIPGKEDL (70)
UBP8_SC SGLINMGS..TCFMSSILQCLIHNPYFIRHSYSQ (61) LAGYSQQDAHEFWQFLINQLHQSYVLD (14) CECIVHTVFEGSLESSIVCPGCQNNSK.....TTIDPFLDLSLDIKDKKKL (15)
USP22_HS RGLINLGN..TCFMNCIVQALTHTPLLRDFFISD (47) LAGYEQQDAHEFLIAALDVLHRHCKGD (11) CNCIIDQIFTGGLQSDVTCQVCHGVS.....TTIDPFWDISLDLPGSSTP (41)
USP27_HS RGLINLGN..TCFMNCIVQALTHTPILRDFFISD (47) LAGYRQQDAHEFLIAALDVLHRHCKGD (11) CNCIIDQIFTGGLQSDVTCQACHGVS.....TTIDPCWDISLDLPGSCTS (40)
USP51_HS RGLINLGN..TCFMNCIVQALTHIPLLKDFFISD (47) LAGYRQQDAHEFLIATLDVLHRHSKDD (11) CNCIIDQIFTGGLQSDVTCQACHSVS.....TTIDPCWDISLDLPGSCAT (41)
USP16_HS KGLSNLGN..TCFLNAVMQNLSQTPVLRELLKEV (64) FKGYQQQDSQELLRYLLDGVRAEEHQR (32) MPSFVDRIFGGELETSMICDQCRTVS.....LVHESFLDLSLPVLDDQSG (254)
UBP9_SC FGYENFGN..TCYCNSVLQCLYNIPEFRCNVLRY (262) FNTMMQQDAHEFLNFLLNDFSEYIQRN (14) NDNFITDLFKGTLTNRIKCLTCDNIT.....SRDSPFLDFPLEVQGDEET (18)
UBP13_SC FGYENFGN..TCYCNSVLQCLYNLSSLRENILQF (254) FNTTMHQDAHEFFNFLLNELSEYIERE (17) SKNFISDLFGGTLTNQIKCLTCDNIT.....SRDSPFLDFPLEVQGDEET (18)
USP12_HS FGLVNFGN..TCYCNSVLQALYFCRPFREKVLAY (46) FDNYMQQDAHEFLNYLLNTIADILQEE (24) DPTWVHEIFQGTLTNETRCLNCETVS.....SKDSDFLDLSVDVEQNTSI (16)
USP46_HS FGLVNFGN..TCYCNSVLQALYFCRPFRENVLAY (46) FDNYMQQDAHEFLNYLLNTIADILQEE (24) ELTWVHEIFQGTLTNETRCLNCETVS.....SKDSDFLDLSVDVEQNTSI (16)
USP33_HS TGLKNIGN..TCYMNAALQALSNCPPLTQFFLDC (47) FRGYSQQDAQEFLRCLMDLLHELKEQ (180) YRSVISDIFDGTLISSVQCLTCDRVS.....VTLETFQDLSLPIPGKEDL (70)
DOA4_SC VGLENLGN..SCYMNCIIQCLLGTHELTQIFLDD (57) FKTASQQDCQEFCQFLLDGLHEDLNQC (36) DFSVIVDLFGQGQYASRLKCKVCSHTS.....TTYQPFTVLSLPIPKKNSR (21)
UBP5_SC VGLENIGN..CCYMNCILQCLVGTHDLVRMFLDN (60) YSDSMQQDCQEFCQFLLDGLHEDLNQN (36) DFSAILDLFGQGQYASRLQCQVCEHTS.....TTYQTFSVLSVPVPRVKTC (18)
USP8_HS TGLRNLGN..TCYMNSILQCLCNAPHLADYFNRN (52) FAGYSQQDSQELLLFLMDGLHEDLNKA (30) NESIIVALFGGQFKSTVQCLTCHKKS.....RTFEAFMYLSLPLASTSKC (18)
USP36_HS AGLHNLGN..TCFLNATIQCLTYTPPLANYLLSK (47) FRFGNQEDAHEFLRYTIDAVQKACLNG (9) ATTLVHQIFGGYLRSRVKCSVCKSVS.....DTYDPYLDVALEIRQAANI (16)
USP1_HS VGLNNLGN..TCYLNSILQVLYFCPGFKSGVKHL (79) YEGYLQHDAQEVLQCFLGNIQETCQLL (203) GFELVEKLFQGQLVLRTRCLECESLT.....ERRDDFQDLSLPVQEDELS (35)
UBP15_SC VGFRNQGA..TCYLNSLLQSYFFTKYFKKLVYEI (42) AESFTQHDVQELNRILMDRLENNMKGTP.....VEGKLNEIFVGKLKSYIKCINVDYES.....ARVDFWDLQLNVKNFKNL (16)
USP7_HS VGLKNQGA..TCYMNSLLQTLFFTNQLRKAVYMM (42) LDSFMQHDVQELCRVLLDNVENKMKGTC.....VEGTIPKLFRGKLVSYIQCKEVDYRS.....DRRDYYDIQLSIKGKKNI (16)
USP54_HS KGLSNEPGQNSCFLNSALQVLWHLDIFRRSFRQL (48) FQLGIMDDAAECFENLLMRIHFHLADE (8) QHCIISHQKFAMTLFEQCVCTSCGATSDPL....PFIQMVHYISTTSLSNQAI (22)
USP35_HS IGLINLGN..TCYVNSILQALFMASEFRHCVLRL (40) FSPGTQQDCSEYLKYLLDRLHEEKTG (22) SSTSVEKMFGGKLVIRICCLCLNWESRE.....SREAFTDLSLAFPPPERC (171)
USP30_HS PGLVNLGN..TCFMNSLLQGLSACPAFIRWLEEF (54) ISSFEEQDAHELFHVITSSLEERDRQ (35) NHWKSQHPFGRLTSNVVCKHCEHQSP.....VRFDTFDSLSLSIPAATWG (18)
USP18_HS VGLHNIGC..TCCLNSLIQVFVMNVDFTRILKRI (45) VPLFVQHDAAQLYLKLWNLIKQLTDVH.....LVERQALYTIRVKDSIICVDCAMES.....SRNSSMLTLPLSLFDVDSK (21)
USP32_HS TGLSNLGN..TCFMNSSIQCLSNTQPLTQYFISG (52) FNGFQQQDSQELLAFLLDGLHEDLNRV (30) NRSIVVDLFHGQLRSQVKCKTCGHIS.....VRFDPFNFLSLPLPMDSYM (332)
UBP10_SC RGLLNHGV..TCYINAAVQALVHIPSIQHYIFDI (57) MSEWQQEDSHEYFMSLWSRLQEDSVPKGHKL...IESIIVDIFGGLLIKQIVTCKSCGSIS.....KTEQPFYDLSLHIKGKKKL (79)
USP34_HS VGLTNLGA..TCYLASTIQQLYMIPEAFQAVFTA (45) LNTGEQKDMTEFFTDLITKIELMSPEL.....KNTVKSLFGGVITNNVVSLDCEHVS.....QTAEEFYTVRCQVADMKNI (16)
USP38_HS TGLINLGN..TCYMNSVIQALFMATDFRQVLSL (40) FTPRSQQDCSEYLRFLLDRLHEEKIL (38) EKTLIEKMFGGKLRTHFRGLNCRSTS.....QKVEAFTDLSLAFCPSSSL (108)
USP45_HS RGIINLGN..TCFFNAVMQNLAQTYTLTDLVNEI (64) FKDFQQQDSQELLHYLLDAVRTEETKR (33) KMNFIDRIFIGELTSTVMCEECANIS.....TVKIPFIDILSLPIIEERVS (246)
USP17L_HS AGLQNMGN..TCYENASLQCLTYTLPLANYMLSR (41) FHRGKQEDVHEFLMFTVDAVKKACLPG (9) DTTLIHQIFGGCWRSQIKCLHCHGIS.....DTFIPYLDLALDIQAAQSV (16)
UBP6_SC VGFKNMGN..TCYLNATLQALYRVNDLFDMILNY (61) GGFYKQQDAEELFTQLFHSVSIVFGDK..........FSEDFRIQFKTTIKDTANDNDITV.....KENESDSKLQCHISGTTNFMRNGL..
USP14_HS CGLTNLGN..TCYMNATVQCLRSVPELKDALKRY (55) QGQYLQQDAQEVCWIQMMRVLQQKLEAI (20) KKSLIDQFFGVEFETTMKCTESEEEEVT.....KGKENQLQLSCFINQEVKYLFTGLK.
USP17_HS AGLQNMGN..TCYVNASLQCLTYTPPLANYMLSR (41) FHRGKQEDAHEFLMFTVDAVKKACLPG (9) DTTLIHQIFGGCYWRSQIKCLHCHGIS.....DTFIPYLDLALDIQAAQSV (16)
UBP12_SC TGLVNLGN..TCYMNSALQCLVHIPQLFDYFLYD (58) FSGYMQQDSQEFLAFLLDSLHEDLNRI (37) NCSVITDLFVGMYKSTLVCPECQNVS.....ITFIPYNDVTLPIPVDTVW (418)
USP4_HS CGLGNLGN..TCFMNSALQCLSNTAPLTDYFLKD (52) FSGYQQQDSQELLAFLLDGLHEDLNRV (30) NDSVIVDTFHGLFKSTLVCPECAKVS.....VTFIPFCYLTLPLPLKKDR (307)
USP11_HS CGLTNLGN..TCFMNSALQCLSNVPQLTEYFLNN (52) FLGYQQHDSQELLSFLLDGLHEDLNRV (30) NDSVIVDTFHGLFKSTLVCPDCGNVS.....VTFIPFCYLSVPIPISHKR (303)
USP15_HS CGLSNLGN..TCFMNSAIQCLSNTPPLTEYFLND (52) FSGYQQQDCQELLAFLLDGLHEDLNRI (30) NDSIIVDIFHGLFKSTLVCPECAKIS.....VTFIPFCYLTLPLPMKKER (330)
USP9X_HS VGLKNAGA..TCYMNSVIQQLYMIPSIRNGTLAAI (89) VNLREQHDALEFFNSLVDSIDEALKALG.....HPAMLSKVLGGSFADQKICQGCPHRY.....ECESSFTTLNVDIRNHQNL (16)
USP41_HS VGLHNIGC..TCCLNSLIQVFVMNVDFARILKRI (45) VPLFVQHDAAQLYLKLWNLIKQLADVH.....LVERQALYMIRVKDSIICLDCAMES.....SRNSSMLTLRLSFFDVDSK (21)
PAN2_SC SGLDPDVD..NHYTNAIIQLYRFIPEWFNFVVGC (74) SESIKRNMPQKFNRFLISQLIKEEAQTVN.....HNITLNQCFGLETEIRTECLSCDHYD.....TTVKLLPSLSISGINKTVI (25)
USP52_HS AGLEPHIP..NAYCNCMIQVLYFLEPVRCLLQNH (56) GKGNLARLIQRVNRFLLTQLHQDVQEL (17) GDSVIGQLFSCEVNCSLCLRCGSETVRASSTLLFTLSYPDDKTG (18)
UBP7_SC TGLRNLGN..TCYINSMIQCLFAAKTFRTLFTSS (55) KIPDDQQDTQEFLMILLDRLHQELSDQ (36) GISPLDDIFGGQLENSLQCKRCGYTLFNYSTFYVLSLAIPRRSMK (30)
SAD1_SC IGFTNAAT..YDYAHSVLLLISHMVPVRDHFLLN (42) GLNLNPIDPRLFLLWLFNKICSSSNDLKSIL...NHSCKGKVKIAKVENKPEASESVTGK.....VIVKPFWVLTLDLPEFSPF (14)
USP39_HS VGINNIKA..NDYANAVLQALSNVPPLRNYFLEE (54) FQITKQGDCVDFLSWFLNALHSAIGGTKKK...KKTIVTDVFQGSLRIFTKKLPHPDLP (16) MVESTFMYLTLDLPTAPFY (21)
UBP1_SC AGLVNDGN..TCFMNSVLQSLASSRELMEFLDNN (106) LLGYDQEDAQEFFQNLLAELESNVKSL (53) TPFKLMTPLDGITAERQGCLQCGENGG.....IRYSVFSGLSLNLPNENIG (18)
UBP3_SC RGIINRAN..ICFNSSVLQVLYCKPFIDVINVL (86) LQWGHQEDAEEFLTHLLDQLHEEISA (88) VPSPILSKLFGGQFRSVIDIPNNKESQS.....ITLDPFQTIQLDISDAGVN (9)
USP10_HS RGLINKGN..WCYINATLQALVACPPVYHLKFI (69) SEKGRQEDAEEYLGFLLNGLHEELNL (55) VQTPITGIFGGHFRSVVYQ.QSSKES.....ATLQPFFTLQLDLQSDKIR (14)
USP49_HS TGLRNLGN..TCYMNSILQVLSHLQKFRECFLNL (107) FRGYDQQDAQEFLCELLHKVQQELESE (18) VLKVVNTIFHGQLLSQVTCISCNYKS.....NTIEPFWDLSLEFPERYHC (30)
UBP16_SC VGLINRGN..DCFITSSLQGLAGIPRFVEYLKRI (77) KISSKQNDAHEFTLILLQTLQERSKL (8) NLNIPKFPFEGETSKFLVCLKCKGLSE.....PSYKQTFIRELSVPQQTSE (14)
USP31_HS VGLTNLGA..TCYVNTFLQVWFLNLELRQALYLC (53) LDTGQQQDAQEFSKLFWSLIEDTLSKQKNPD...VRNIVQQQFGGEYAYVTVCNQCGRES.....KLLSKPYELELNIQGHKQL (16)
USP29_HS QGFPNLGN..TCYMNAVLQSLFAIPSFADDLLTQ (48) FSGNMQNDAHEFLGQCLDQLKEDVEKL (28) FVCPVVANFFEIQLSLICKACGHAV.....LKVEPNNYSLTNIHQETKP (17)
UBP2_SC TGINNIGN..TCYLNSLLQYFSIAPLRRYVLEY (165) LEMGRQQDVTECLGNVLFQLESGSEPI (7) QYDLVKQLFYCTTKQSIVPLSATNKVR.....TKVERFLSLLINIGDHPKDIYDAF..
USP25_HS VGLKNVGN..TCWFSAVIQSLFNLLEFRRLVLNY (54) SNDSQQQDVSEFTHKLLDWLEDAFQMK (9) PKNPVVELFYGRFLAVGVLEGKKFEN (8) LQVNGFKDLHECLEAAMIE.....
USP28_HS VGLKNVGN..TCWFSAVIQSLFQLPEFRRLVLSY (54) SSEEQQQDVSEFTHKLLDWLEDAFQLA (9) SENPVVQLFYCTFLTEGVREGKPFCN (8) LQVNGYRNLDECLEGAMVE.....
USP3_HS TGLRNLGN..TCFMNAILQSLSNIEQFCCYFKEL (62) FRGYQQQDAHEFNALPFGPPTLGNFRA (27) ASTVVTAIFGGILQNEVNCLICGTES.....RKFDPFLDLSLDIPSQFRS (30)
USP26_HS HGLPNLGN..TCYMNAVLQSLLSIPSFADDLLNQ (48) FHGNAQNDAHEFLAHCLDQLKQNTEKL (28) FSCPVITNFELELLHSICKACKACGQVI.....LKTELNNYLSTNLPQRIKA (17)
USP37_HS QGFSNLGN..TCYMNAILQSLFSLQSFANDLLKQ (48) FSGYMQNDAHEFLSQCLDQLKEDVEKL (23) YTCPVITNLEFEVQHSICKACGEII.....PKRFQFNDLSLDLPRRKKP (18)
UBP11_SC TGLQNPGN..TCYINSIIQCLFGTTLFRDLFLTK (52) NIPDDQQDTQEFLLIVLARIHEELSNE (35) GLSPLDHIYRGQLENILKCQRCGNSS.....YSYSTFYVLSLAIPKLSLY (27)
USP2_HS AGLRNLGN..TCFMNSILQCLSNTRELRDYCLQR (51) FVGYNQQDAQEFLRFLLDGLHNEVNRV (30) EDSRIGDLFVGQLKSSLTCTDCGYCS.....TVFDPFWDLSLPIAKRGYP (20)


|← % %
box 4 →|
|← box 5 * →|
|← *? box 6 →|
USP24_HS
USP9Y_HS
UBP14_SC
USP5_HS
USP13_HS
USP21_HS
USP19_HS
USP20_HS
UBP8_SC
USP22_HS
USP27_HS
USP51_HS
USP16_HS
UBP9_SC
UBP13_SC
USP12_HS
USP46_HS
USP33_HS
DOA4_SC
UBP5_SC
USP8_HS
USP36_HS
USP1_HS
UBP15_SC
USP7_HS
USP54_HS
USP35_HS
USP30_HS
USP18_HS
USP32_HS
UBP10_SC
USP34_HS
USP38_HS
USP45_HS
USP17L_HS
UBP6_SC
USP14_HS
USP17_HS
UBP12_SC
USP4_HS
USP11_HS
USP15_HS
USP9X_HS
USP41_HS
PAN2_SC
USP52_HS
UBP7_SC
SAD1_SC
USP39_HS
UBP1_SC
UBP3_SC
USP10_HS
USP49_HS
UBP16_SC
USP31_HS
USP29_HS
UBP2_SC
USP25_HS
USP28_HS
USP3_HS
USP26_HS
USP37_HS
UBP11_SC
USP2_HS

3.5.2.3 Large insertions between the individual boxes

The distance between the individual boxes varies remarkably between the USPs. The different dimensions of the USP homology domain in different proteins are a direct consequence of variable insertions between the boxes. For example, the boxes in USP7/HAUSP are adjacent or have only short insertions in between, while USP1 has a ~200 residue insert between boxes 3 and 4 and another large ~150 residue insert between box 5 and box 6 (see Figure 3-11). Overall, the USP domain in USP1 spans ~700 residues as opposed to ~300 in USP7. Nonetheless, the generalized profile constructed from an initial set of USP proteins was flexible enough to report the proper USP domain even in USP1 and other extreme cases, which had not been part of the training set.

Another prominent example for a rather large insertion is USP15 with ~330 residues between boxes 3 and 4. Interestingly, the region following the C-terminus of box 3 shares homology with a domain of USP7, which here is located C-terminally of box 6. This domain was detected during a homology domain analysis of USP7/HAUSP carried out in this work and has the preliminary working title HAUSP2A. It has no known function or structure (not part of the crystallized USP7), but it is detectable in several USP family members besides USP15, suggesting a role in deubiquitylation. Human USP13 is an example for a large insertion between box 4 and box 5. Here, two UBA domains are inserted (data not shown). Likewise, three UIMs are inserted between the same boxes in human USP37. As the boxes of the USP domain can accommodate even the insertion of large or multiple homology domains, the question arises how the catalytic domain fold still can be stabilized.

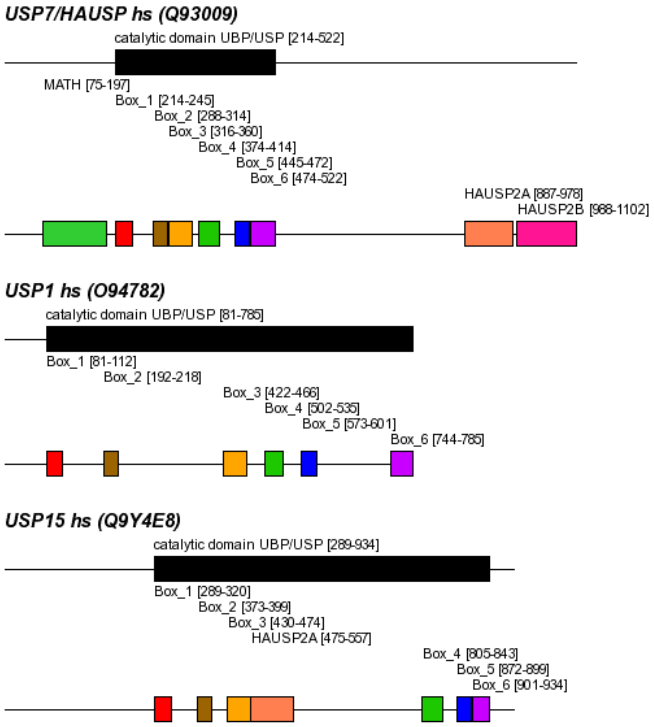


Figure 3-11 Domain structure of selected human USPs. In addition to the catalytic domain (black box) the positions of the six blocks (named Box_1-6) that constitute the catalytic domain are depicted. The regions between these blocks are regarded as insertions not conserved among the USP family. USP7 has a so-called HAUSP2A domain located near the C-terminus. The same domain is found between box 3 and box4 in USP15. Note the different dimensions of the catalytic domain in USP7 and USP1.

3.5.2.4 A potential zinc finger is inserted into the catalytic domain

Note: some of the following results have been published in Current Biology (Hetfeld, 2005).

Regarding the question of how the USP fold is maintained despite large insertions, the observation of two conserved cysteine dyads in boxes 3 and 4 is of special interest. There is evidence that these two dyads work together as their occurrence in the two boxes is found to be highly correlated. The cysteines within each dyad are in nearly all cases separated by two residues, reminiscent of classic C4 zinc fingers. To find further evidence for a putative zinc finger behaviour, the only available USP structure, which is USP7/HAUSP, was analysed. Unfortunately, USP7 is one of the very few USPs lacking the dyads and only a single cysteine appears retained. Indeed, no zinc ion is visible in the part of the structure that corresponds to the putative zinc finger. However, inspection of this region in USP7 revealed a secondary structure arrangement similar to that in known zinc fingers (see Figure 3-12). There are two β-hairpins that lie in a perpendicular orientation. The four residues at positions homologous to the cysteine dyads in other USPs occupy positions in the turns of the β-hairpins. Therefore, this region of USP7 adopts a fold analogous to classic zinc ribbon structures, though no zinc is coordinated. A

similar observation has been discussed in chapter 3.4 for a zinc-less derivative of the RING finger, the U-box domain, which shows that the zinc-less zinc finger fold of USP7 is not a structural singularity. It should be mentioned that the crystal structure of the USP member yeast Ubp6 has recently been solved, but this USP lacks both cysteine dyads and the region corresponding to the zinc finger is disordered (Murshudov, 2004, unpublished, submitted to PDB as 1VJV).

The zinc finger in USPs may have a structural role and stabilize the tertiary structure, when there are large insertions between box 3 and 4. Indeed, most USPs without the cysteine dyads and therefore without the capability to coordinate zinc show relatively small insertions between boxes 3 and 4.

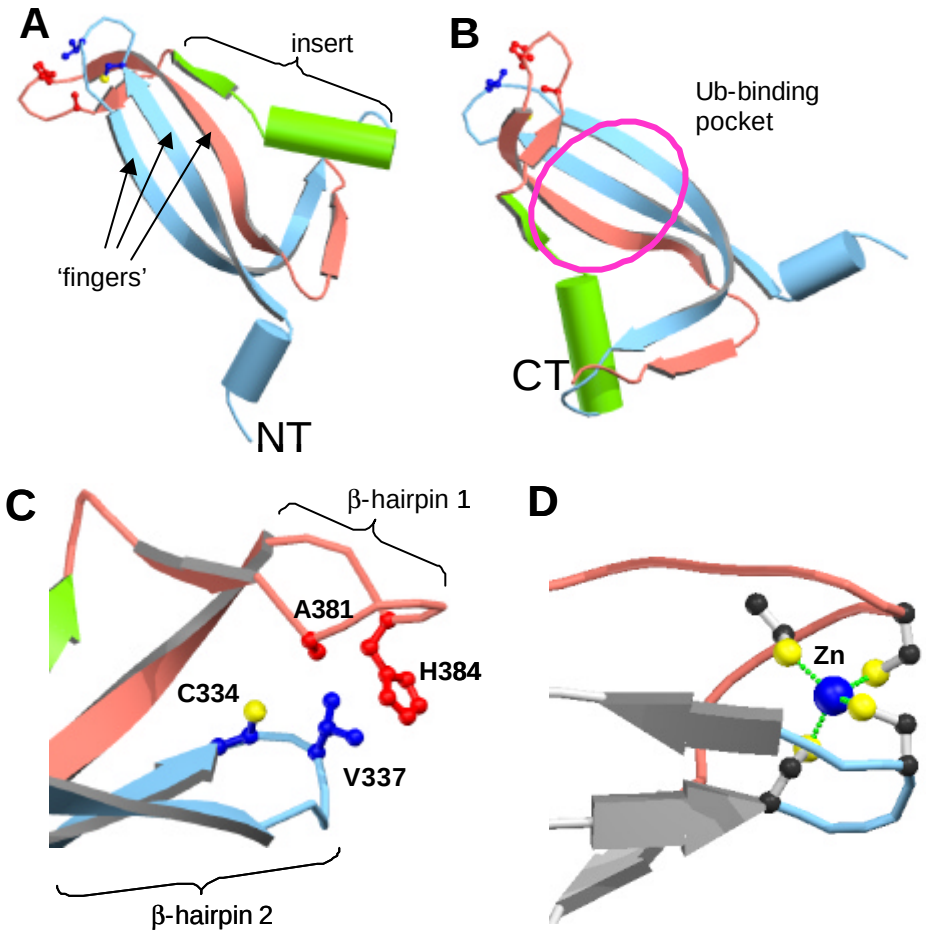


Figure 3-12 Zinc finger like fold between box 3 and box 4 in USP7. (A), section of USP7 comprising box 3 (blue) and box 4 (red) (pdb:1NBF, residues 316 to 414). The insert between these two boxes is coloured in green and part of a larger structural domain ('palm') not shown. (B), a purple circle indicates the Ub-binding pocket. (C), close-up view on the region homologous to the putative zinc finger in other USPs. Side chains of residues homologous to cysteines in the USP dyads are shown in ball-and-stick representation. These residues are distributed over two β-hairpins. Yellow indicates sulphur atoms, while other atoms are coloured according to the box to which they belong. (D), zinc finger of Rubredoxin (pdb:1H7V), which was detected by profiles derived from boxes 3 and 4. The blue sphere indicates a zinc ion. The two hairpins harbouring the coordinating cysteines are coloured in red and blue, respectively.

3.5.3 Other families

3.5.3.1 A cysteine-protease motif in ataxin-3

Note: the following work has been published as "Elucidation of ataxin-3 and ataxin-7 function by integrative bioinformatics." by Scheel, H. et al. in Hum Mol Genet 12(21): 2845-52 (Scheel et al., 2003).

The ataxin-3 (MJD1) gene is mutated in spinocerebellar ataxia type 3 (SCA3) (Kawaguchi, 1994). The corresponding protein harbours an uninterrupted stretch of multiple glutamine residues, which is considerably longer in the disease state as a consequence of the pathological expansion of an unstable (CAG)n triplet repeat found in the coding portion of the SCA genes. The expanded poly(Q) region is thought to contribute crucially to SCA3 pathogenesis, as ataxin-3 induces the formation of intranuclear inclusion bodies. The mutated protein is found within these inclusions, typically associated with other components such as Ubiquitin and chaperones. Despite a lot of research into this topic, the pathological mechanism of this disease is not entirely clear and several pathogenesis models exist, which are not mutually exclusive. Poly(Q) inclusions are known to sequester other glutamine-rich proteins, which might be required in soluble form for cell viability (Gusella, 2000, Margolis, 2001, Uchihara, 2001). In model organisms, poly(Q) toxicology can be rescued by overexpression of chaperones, suggesting an involvement of the cellular quality control mechanism in pathogenicity, an idea supported by the presence of Ubiquitin and proteasomal components in the inclusion bodies. Finally, a loss-of-function of the mutated protein might contribute substantially to the disease. In most cases, this latter contribution cannot really be assessed, as the function of many ataxins is not known.

Ataxin-3 has been found to bear three UIM copies in the C-terminal portion of the protein, an observation originally reported by Hofmann et al. (Hofmann et al., 2001). As the UIM in general appears to be closely connected to pathways of protein ubiquitylation and Ub-recognition, ataxin-3 is suggested to work in the UPS. In the course of this work it became obvious that the human genome encodes several ataxin-3-like proteins and that the N-terminal part of ataxin-3 is the most highly conserved portion of the protein. Therefore, sequence analysis was focussed on the functional role of this protein domain.

In the first step, current versions of the protein database were searched for proteins with similarity to the ataxin-3 N-terminal domain. In order to obtain the full complement of mammalian ataxin-3-like proteins, the EnsEMBL version of the human genome database and translated versions of the current mammalian EST databases were also searched (Hubbard et al., 2002). As shown in Figure 3-13, it was possible to identify four ataxin-3-like proteins in

mammals. Besides ataxin-3, there is one closely related protein (Atx3like) sharing the same domain architecture comprising the conserved N-terminal domain followed by multiple UIMs and a poly(Q) stretch. Two shorter proteins that have been published before as Josephin-1 and Josephin-2 consist of not much more than the N-terminal domain (Albrecht, 2003). Nematodes and plants seem to possess two ataxin-3-like proteins, one of them corresponding to the long form, the other to the short Josephin-form, while insects only have a copy of the latter. There is no evidence for ataxin-3- like proteins in yeasts, but there are two proteins of that class found in the parasite *Plasmodium falciparum* (see Figure 3-14). In the longer plasmodium protein, the conserved ataxin-3 N-terminal domain is followed by a Ub-like UBX domain instead of the UIM motifs (see Figure 3-15). In the hope of finding more distantly related members of this family with some degree of functional annotation, generalized profiles were constructed from the multiple alignment shown in Figure 3-13. Unfortunately, no related proteins could be identified in this screen, using the stringent criterion of $P<0.01$ (data not shown).

However, a useful similarity information on the ataxin-3 N-terminal domain family was serendipitously obtained by using a completely different screen that was originally targeted for identifying new activating proteases for Ub and related modifiers. After exhaustive application of conventional generalized profile searches starting from multiple alignments of the UCH protease family, the profile searches were made more sensitive to very distantly related outlier sequences by incorporating information from three-dimensional structures into the profile construction process. According to the FSSP database, which holds information on structural relationships calculated by the Dali algorithm, the closest known structural neighbour of the UCHL3 protease (PDB:1UCH) is a leader protease (PDB:1QMY) from the foot and mouth disease virus (Holm, 1997). The corresponding sequences do not show any recognizable similarity, although the general fold and the active site geometry of the two enzyme classes are clearly related. The spdbv program was used to calculate a rigid body superposition of the two structures (Guex et al., 1997). From this superposition, a structurally valid two-sequence alignment was derived; a number of manual adjustments were required to overcome the rigid-body limitations. After incorporation of other UCH proteases and members of the viral protease class, generalized profile searches were performed starting from this alignment. In this database search, several high-significance matches to members of the USP family of Ubiquitin-specific proteases were obtained. The significant sequence relationship between the UCH and USP classes of ubiquitin proteases was somewhat surprising. However, the published structure of HAUSP, the first structurally characterized USP-type protease, clearly demonstrates the structural relationship and active site correspondence between these two protease classes (Hu et

al., 2002). Thus, the significant profile scores obtained in the screen are biologically meaningful and underscore the suitability of the chosen approach.

The structure-based profile searches were further refined by also including members of the USP family into the training set. The resulting profile, consisting of members of the three protease classes, found significant matches to several members of the ataxin-3 family described above. While not all of the structural features of the UCH/leader protease/USP profile can be reliably mapped to the ataxin-3 sequence, the catalytically most important regions can be aligned with high confidence (see Figure 3-16). The catalytic cysteine residue corresponds to C14 of human ataxin-3, while the proton-donating histidine residue corresponds to the H119 position. As the third residue of the catalytic triad, an aspartate is observed in all UCH and viral proteases, while in USP proteases either aspartate or asparagine can be found in that position. Members of the ataxin-3 family share this variability: ataxin-3 itself carries an asparagine (N134) while most other ataxin-3 like proteins use an aspartate instead. Based on the statistically significant profile comparison scores and the conservation of the catalytically important residues, it is predicted that all members of the ataxin-3 family are cysteine proteases assuming the same fold as UCH, USP and FMDV leader proteases, i.e. the papain fold (Anantharaman, 2003).

The USP-type proteases contain a well-conserved asparagine or glutamine residue shortly upstream of the catalytic cysteine, which balances the oxyanion hole of the transition state, probably in a concerted manner with another hydrogen bond providing group from either a backbone imino- or side chain amino-group (Hu et al., 2002); this feature is also shared by papain and most other proteases belonging to this fold (Anantharaman et al., 2003). Ataxin-3 and its relatives carry a highly conserved glutamine residue at the corresponding position, which is likely to assist in the catalytic reaction. The substrate specificity of ataxin-3 cannot be derived from this kind of bioinformatical analysis. However, the presence of a UIM domain in ataxin-3 and the presence of a UBX domain in the *P. falciparum* homologue suggest a role of this protein family in ubiquitin-dependent pathways.

Based on the results from sequence analysis done here, ataxin-3 is predicted to be a cysteine protease assuming the papain fold that is structurally and functionally related to two quite divergent subfamilies of Ub-specific proteases, the UCH and the USP enzymes. Indeed, Burnett et al., and later Chow et al., confirmed this prediction soon after its publication (Burnett et al., 2003, Chow, 2004).

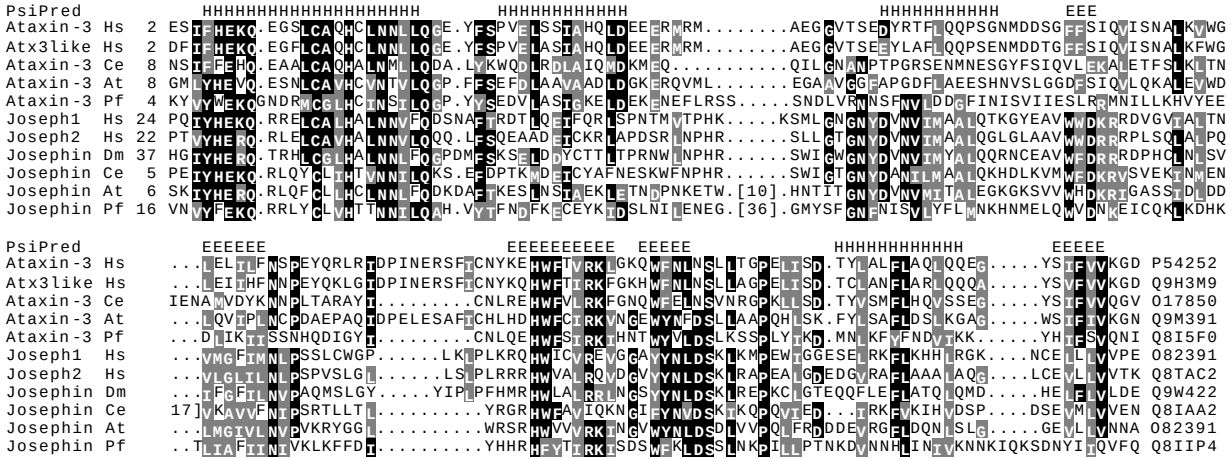


Figure 3-13 Multiple alignment of the ataxin-3 N-terminal domain. Positions invariant or conservatively replaced in at least 50% of the sequences are shown on black and grey background, respectively. The first two columns indicate the gene name and the species abbreviation (Hs, *Homo sapiens*; Ce, *Caenorhabditis elegans*; At, *Arabidopsis thaliana*; Dm, *Drosophila melanogaster*; Pf, *Plasmodium falciparum*). The SwissProt/TrEMBL accession numbers of the sequences are shown in the last column. The top line contains the PsiPred secondary structure prediction for ataxin-3, H denoting α-helices and E extended/sheet structures. The PredictProtein/PROF prediction largely agrees with PsiPred and is not shown.

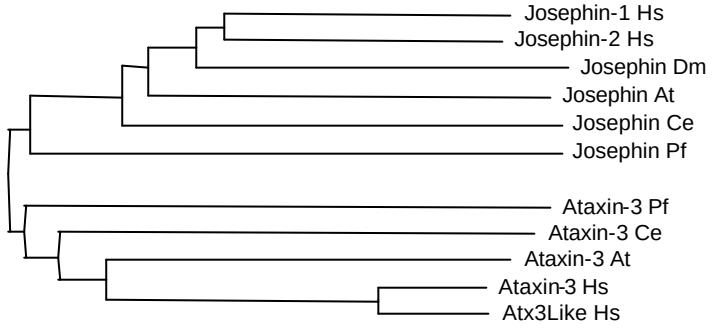


Figure 3-14 Evolutionary relationship of the ataxin-3/Josephin family, as determined by neighbour-joining dendrogram analysis.

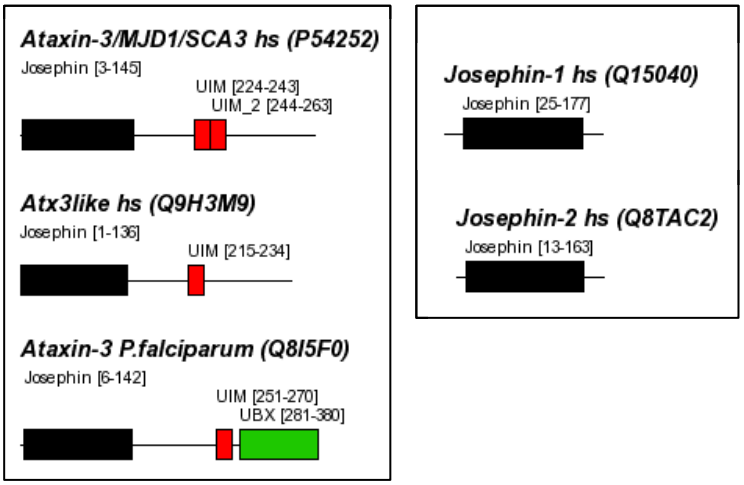


Figure 3-15 Domain organisation of representative ataxin-3 related proteins. The N-terminal box labelled 'Josephin' refers to the putative protease domain discussed in the text. The red-coloured boxes in the C-terminal region of ataxin-3 denote multiple copies of the UIM motif. A third copy is only present in selected splice forms of ataxin-3 and is not shown here. The box labelled 'UBX' denotes a Ub-fold UBX domain in the C-terminus of the Plasmodium version of ataxin-3.

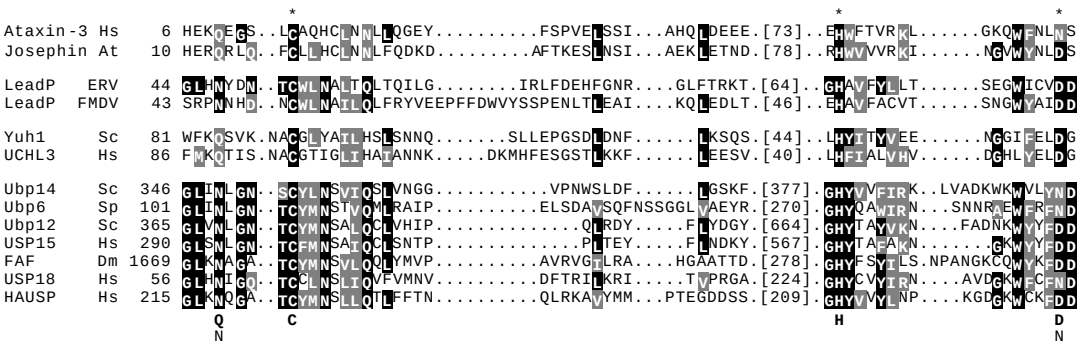


Figure 3-16 Multiple alignment of representative members of the protease families discussed in the manuscript. Shading of conserved residues and species abbreviations are analogous to Figure 3-13. The topmost block contains human ataxin-3 and a plant Josephin sequence; the second block contains leader proteases from foot and mouth disease virus (FMDV) and equine rhinitis virus (ERV). The third block contains two UCH type proteases (yeast Yuh1 and human UCHL3). The bottom block contains a divergent set of USP-type proteases. Residues important for catalysis are indicated by their consensus symbols below the alignment.

3.5.3.2 OTU family

The OTU (ovarian tumour) proteases are an additional class of cysteine proteases with a deubiquitylating activity, which has been shown for several members of this family (Evans et al., 2003, Soares et al., 2004). The OTU family is characterized by a catalytic domain that contains the same catalytic triad as found in the UCH and USP family (Makarova et al., 2000). Profiles constructed from this domain detected two yeast proteins and in addition 14 human proteins (see Table 3-23).

Table 3-23 Yeast and human OTU proteins. A subtable contains orthology assignments.

Yeast		Human			
Gene name	ORF	Gene name	Uniprot	Gene name	Uniprot
YOD1	YFL044C	OTUB1	Q96FW1	VCIP135	Q96JH7
YHL013C	YHL013C	OTUB2	Q96DC9	FLJ46133	Q6ZRS6
Assignment of orthologues		ZA20D1/Cezanne1	Q6GQQ9	TNFAIP3/A20	P21580
Yeast	**Human**	C15orf16/Cezanne2	Q8TE49	CGI-77	Q8N6M0
YOD1	FLJ46133	HIN1	Q01804	FLJ25831	Q7L8S5
YHL013C	CGI-77, FLJ25831	hin1L	Q7RTX8	DKFZp761A0	Q96G74
		ZRANB1	Q9UGI0	OTDC1	Q5VV17

3.5.3.3 MPN family

Another family of proteases for Ub and Ub-related modifiers is the MPN family. This family differs from the DUBs described so far in that its activity depends on bound metal ions instead of a cysteine-based catalytic triad found in other proteases. According to sequence analysis, the MPN family may be subdivided into two groups, which differ in the residues of the catalytic site (Maytal-Kivity et al., 2002, Verma et al., 2002). While one subfamily lacks several or all residues needed for zinc coordination and is inactive, the other subfamily (MPN+/JAMM)

is characterized by the presence of all zinc-chelating residues (Maytal-Kivity et al., 2002, Verma et al., 2002).

Yeast encodes four MPN members (see Table 3-24). Of these, Rpn11 and Rri1 belong to the MPN+ subfamily and are active proteases. Rpn11, a known subunit of the lid subcomplex of the 19S regulatory particle, has a deubiquitylating activity (Maytal-Kivity et al., 2002). Rri1 (or Csn5), which is a component of the lid-related COP9 signalosome complex (CSN), has been assigned a role in cleaving Rub1, the yeast orthologue of NEDD8, from Cdc53, the cullin subunit of the SCF complex (Cope et al., 2002).

The human orthologues of yeast Rpn11 and Rri1 occupy analogous positions in the corresponding human lid and CSN complexes and like their yeast counterparts, are members of the MPN+/JAMM subfamily. A third human member of this subfamily is STAMBP (or AMSH), for which a deubiquitylating activity has been shown recently (McCullough, 2004). Like STAMBP, the remaining human MPN members that are not subunits of the lid, the CSN or the eIF3 complexes, are all catalytically active MPN+/JAMM members.

Table 3-24 Yeast and human MPN proteins. A subtable contains orthology assignments. Members of the MPN+/JAMM subfamily and therefore active or potentially active MPNs are underlined. Genes in italics indicate pseudogenes.

Yeast		Human	
Gene name	ORF	Gene name	Uniprot number
RPN11	YFR004W	C6.1A	P46736
RPN8	YOR261C	CSN5/COPS5	Q92905
RRI1	YDL216C	CSN6/COPS6	Q7L5N1
PRP8	YHR165C	eIF3h/EIF3S3/p40	O15372
Assignment of orthologues		eIF3f/EIF3S5/p47	O00303
Yeast	**Human**	PSMD7	P51665
PRP8	hPRP8	PSMD14	O00487
RPN11	PSMD14	hPRP8	O14547
RPN8	PSMD7	STAMBP	O95630
RRI1	COPS5	FLJ14981	Q8N594
		AMSH-LP	Q96FJ0
		KIAA1915	Q96PX3
		IFP38	*Q9BX72*

3.5.3.4 Desumoylating enzymes

Proteases that process SUMO-precursors and SUMO-conjugates do not belong to any of the DUB families described above. Instead, they constitute an ancient family present in fungi and mammals known as ULP (Ub-like protein specific protease). An evolutionary relationship to adenoviral proteases has been proposed (Yeh, 2000) and could be confirmed in this work. The catalytic domain was chosen as a suitable starting point for profile construction and searching for other family members. As a result, two ULP members were found in yeast opposed to seven in human (see Table 3-25). The evolutionary relationship between the yeast and human ULPs was somewhat difficult to assign. No clear 1:1 assignments could be made, for example yeast Ulp2 was classified as orthologue of human SENP6 and SENP7 in this work. At

the same time, dendrogram analysis carried out here and available functional data on Ulp1 reported by Li et al. (Li, 2003) suggest an orthologous relationship to SENP1 and SENP2. It should be noted that SENP8/Den1 has also been reported to be active as deneddylating enzyme (Gan-Erdene, 2003). Thus, Nedd8/Rub1 seems to employ members of two distinct protease classes, the initially reported MPN family (Csn5) and the ULP family.

Table 3-25 Yeast and human ULP proteins.

Yeast		Human	
Gene name	ORF	Gene name	Uniprot number
Ulp1	YPL020C	SENP1	Q9P0U3
		SENP2	Q9HC62
-	-	SENP3	Q9H4L4
-	-	SENP5	Q96HI0
Ulp2	YIL031W	SENP6	Q9GZR1
		SENP7	Q9BQF6
-	-	SENP8/Den1	Q96LD8

3.6 Ub- and SUMO-binding proteins

3.6.1 The UBA domain and its relatives

3.6.1.1 The UBA domain family comprises several subfamilies

The UBA domain, the first described general Ub-binding domain, can be subdivided into at least six different subfamilies with the 'classic UBA domain' forming by far the largest group (Table 3-26). Several of the other subfamilies have been described as autonomous Ub-binding domains in the past, but, on closer look, are divergent variants of the classic UBA domain.

The classic UBA domain was found in a variety of both yeast and human proteins, for example in several Ub conjugating and ligating enzymes, deubiquitylating enzymes and adaptors like Rad23. In addition, UBA domains were also found in proteins that have a role outside ubiquitylation/deubiquitylation and substrate transfer to the proteasome. For example, UBA domains are present in proteins involved in DNA damage repair and endocytosis. A particular frequent domain topology observed is the UBA domain in combination with Ub-like or Ubx domains.

Table 3-26 List of classic UBA proteins in yeast and human as well as orthology assignments. Human orthologues of yeast SNF1 do not contain a detectable UBA domain.

UBA Human Gene name	Uniprot number	Human Gene name	Uniprot number
AD-012	Q9P0H6	RNF31	Q96EP0
C22orf3	Q9Y3P4	SB132	Q96S82
CBL	P22681	SIK2/KIAA0781	Q76N03
CBLB	Q13191	SNF1LK	P57059
DHX57	Q6P158	SNRK	Q6IQ46
EDD	O95071	SQSTM1	Q13501
ETEA	Q9BVM7	STS-1	Q96IG9
FAF1	Q9UNN5	TDRD3	Q9H7E2
GBDR1	O75500	TNRC6C	Q9HCJ0
HERC2	O95714	U33K	Q04323
KIAA0794	O94888	UBAP1	Q6FI75
KIAA0999	Q9Y2K2	UBAP2L	Q14157
LATS1	O95835	UBASH3A	P57075
LATS2	Q9NRM7	UBC1/HIP2	P61086
M17S2	Q14596	UBQLN1/PLIC1	Q9UMX0
MARK1	Q5VTF9	UBQLN2/PLIC2	Q9UHD9
MARK3	P27448	UBQLN3	Q9H347
MARK4	Q96L34	UBQLN4/UBIN	Q9NRR5
NSFL1C	Q9UNZ2	UREB1	Q7Z6Z7
NUB1/NYREN18	Q9Y5A7	USP13	Q92995
OTTHUMP00000000473	Q5T6F2	USP25	Q9UHP3
PHGDHL1	Q8NBM4	USP28	Q96RU2
RAD23A	P54725	USP5/ISOT	P45974
RAD23B	P54727	VPS13D	Q5THJ4

Yeast Gene name	ORF	Orthologues Yeast	Human
DDI1	YER143W	DDI1	DDI1, DDI2
DSK2	YMR276W	RAD23	RAD23A, RAD23B
EDE1	YBL047C	DSK2	UBQLN1-4
GTS1	YGL181W	EDE1	EPS15, EPS15L
RAD23	YEL037C	UBX3	ETEA
RUP1	YOR138C	SHP1	NSFL1C
SEL1	YML013W	SNF1	PRKAA1 (Q13131),
SHP1	YBL058W		PRKAA2 (P54646)
SNF1	YDR477W	UBC1	UBC1/HIP2
SWA2	YDR320C	UBP14	UBP5/ISOT
UBC1	YDR177W	UBX5	KIAA0794
UBP14	YBR058C	YLR419W	DHX57
UBX5	YDR330W		
YLR419W	YLR419W		

3.6.1.2 The CUE domain is related to the UBA domain and may be subdivided

The other 'UBA-related' subfamilies are typically not detected by the classic UBA profile, but reverse profile searches have confirmed their relationship to the classic UBA subfamily. In this work, a relationship between the UBA and the CUE domain could be confirmed by sequence comparison. An alignment of representative CUE domains and UBA domains is shown in Figure 3-17. The UBA domain and CUE domain are highly divergent. Using profiles derived from classic UBA domains, CUE domains were virtually non-detectable. Vice versa, profiles derived from CUE domains were able to retrieve proteins with a classic UBA domain from sequence databases in a significant manner. Taking into account the congruent secondary structure prediction for the UBA and CUE domains, the evolutionary relationship between these two domains is considered real.

Sequence analysis performed in this work suggests that the CUE domain family can be subdivided into two separate subgroups, which are here referred to as CUE-A and CUE-B domains (see Figure 3-18). Interestingly, this classification was corroborated by experimental findings from two groups that have reported the CUE domain of yeast Vps9, Cue2, Cue3, Cue5 and human Tollip to directly bind mono-Ub (Donaldson et al., 2003, Shih et al., 2003). All of these proteins belong to the CUE-B subfamily. By contrast, the CUE domains of yeast Cue1 and Cue5, which are classified as CUE-A type proteins, have been shown to possess only a marginal binding affinity to Ub (Shih et al., 2003).

Table 3-27 CUE proteins. Orthologues are indicated by grey shading.

	Yeast Gene name	ORF	Human Gene name	Uniprot number
CUE-A	CUE1	YMR264W	AMFR	Q9UKV5
	CUE4	YML101C	-	-
	-	-	AUP1	Q9Y679
CUE-B	FUN30	YAL019W	SMARCAD1	Q9H4L7
	CUE2	YKL090W	N4BP2	Q86UW6
	CUE3	YGL110C	-	-
	CUE5	YOR042W	-	-
	DEF1	YKL054C	-	-
	VPS9	YML097C	-	-
	DON1	YDR273W	-	-
	-	-	TOLLIP	Q9H0E2
	-	-	CUEDC1	Q9NWM3
	-	-	DMRT3	Q9NQL9
	-	-	DMRT5	Q96SC8
	-	-	DMRT4	Q5VZB9

```
sec. struc.      HHHHHHHHH   HH  HHHHHHHHHH        HHHHHHHH
JPred           HHHHHHHHH          HHHHHHHHHH    HHHHHHHHH
                   *  ****                       **   *   *
DDI1 SC         TIKQLMDLGF...PRDAVVKALKQTNGN....AEFAASLLFQ.
DSK2 SC         QLRQLNDMGFF..DFDRNVAALRRSGGS....VQGALDSLLNG
EDE1_SC         AVEELSGMGF...TEEEAHNALEKCNWD....LEAATNFLLDS
GTS1 SC         QLAELKDMGFG..DTNKNLDALSSAHGN....INRAIDYLEKS
RAD23 SC 1      TIERIMEMGY...QREEVERALRAAFNN....PDRAVEYLLMG
RAD23 SC 2      AISRLCELGF...ERDLVIQVYFACDKN....EEAAANILFSD
RAD23A HS 1     MLTEIMSMGY...ERERVVAALRASYNN....PHRAVEYLLTG
RAD23A_HS_2     AIERLKALGF...PESLVIQAYFACEKN....ENLAANFLLSQ
RUP1 SC         AVKSLLEMGI...PHEVAVDALQRTGGN....LEAAVNFIFSN     UBA
SEL1 SC         KLNEFQVITNFPPEDLPDVVRLLRNHGWQ...LEPALSRYFDG
SHP1 SC         TIQQFMALTNV...SHNIAVQYLSEFGD....LNEALNSYYAS
UBP14 SC 1      SISQLIEMGF...TQNASVRALFNTGNQD...AESAMNWLFQH
UBP14_SC_2      SLTSMLSMGL...NPNLCRKALILNNGD....VNRSVEWVFNN
YLR419W SC      IVERLTEIGV...SSDEALLALQQNDMN....ENEAAGFLTRE
NUB1 HS 1       KVDNLLQLGF...TAQEARLGLRACDGN....VDHAATHITNR
NUB1 HS 2       NIRFLKGMGY...STHAAQQVLHAASGN....LDEALKILLSN
NUB1_HS_3       NIDRLVYMGF...DALVAEAALRVFRGN....VQLAAQTLAHN

CUE1 SC         MVETVQNLAP....NLHPEQIRYSLENTGS.VEETVERYLRGD
CUE4_SC         MVEIVMTMAP....HVPQEKVVQDLRNTGS.IEHTMENIFAGK     CUE-A
AMFR HS         MAHQIQEMFP....QVPYHLVLQDLQLTRS.VEITTDNILEGR
AUP1_HS         LAQRVKEVLP....HVPLGVIQRDLAKTGC.VDLTITNLLEGA

VPS9 SC         TLNTLQNMFP....DMDPSLIEDVCIAKKSRIGPCVDALLSLS
DON1_SC         VLKELKIAFP....EVDDTLIKAILIASQGVLEPAFNSLLYYS
CUE2 SC 1       KLSILMDMFP....AISKSKLQVHLLENNNDLDLTIGLLKEN
CUE2 SC 2       ELHQLYDMFP....QLDCSVIKDQFVINEKSVESTISDLLNYE
CUE3 SC         QLSALMELFP....QFSKYQLSQTLLAYDNNIELVTNKIFEDP
CUE5 SC         ILQELKDAFP....NLEEKYIKAVIIASQGVLSPAFNALLFLS
DEF1_SC         KIDTLTELFP....DWTSDDLIDIVQEYD.DLETIIDKITSGA     CUE-B
FUN30 SC        ALVNLAREFP....DFSQTLVQAVFKSNSFNLQSARERLTRLR
DMRTA2 HS       PLDILTRVFP....GHRRGVLELVLQGCGGDVVQAIEQVLN..
TOLLIP HS       DLKAIQDMFP....NMDQEVIRSVLEAQRGNKDAAINSLLQMG
SMARCAD1 HS 1   KLQTLKELFP....QRSDNDLLKLIESTSTMDGAIAAALLMFG
SMARCAD1_HS_2   IVLKLQKEFP....NFDKQELREVLKEHEWMYTEALESLKVFA
CUEDC1_HS       AMDDFKTMFP....NMDYDIIECVLRANSGAVDATIDQLLQMN

ARIH1 HS        MVECIREVNEV..IQNPATITRILLSHF....NWDKEKLMERY
ARIH2 HS        LNEHMTSLASV..LKVSHSVAKLILVNF....HWQVSEILDRY
PARC HS         MKQTVRQVQET..LNLEPDVAQHLLAHS....HWGAEQLLQSY     AriNT
ANKIB1 HS      KDMLIVETADM..LQAPLFTAEALLRAH....DWDREKLLEAW
YKR017C_SC      MLQRVDHLQPI..FAIPSADILILLQHY....DWNEERLLEVW

TTRAP HS        RLLCVEFASV...ASCDAAVAQCFLAEN....DWEMERALNSY
ZA20D1 HS       DAVLSDFVRS...TGAEPGLARDLLEGK....NWDVNAALSDF
Q6PH85 HS       KDKVRQFMAC...TQAGERTAIYCLTQN....EWRLDEATDSF
MGC29814_HS     QVMINQFVLA...AGCAADQAKQLLQAA....HWQFETALSTF
LOC124402 HS    QVMINQFVLT...AGCAADQAKQLLQAA....HWQFETALSAF     TtrapNT
C15orf16 HS     DAVLSDFVRS...TGAEPGLARDLLEGK....NWDLTAALSDY
DCN1 SC         QEAIESFTSL...TKCDPKVSRKYLQRN....HWNINYALNDY
RP42 HS         KDKVRQFMIF...TQSSEKTAVSCLSQN....DWKLDVATDNF
NSFL1C_HS       QEALREFVAV...TGAEEDRARFFLESA....GWDLQIALASF

EGD2 SC         NKDDIELVVQQ..TNVSKNQAIKALKAH....NGDLVNAIMSL
KIAA0363 HS     ELRDIELVMAQ..ANVSRAKAVRALRDN....HSDIVNAIMEL     NACαCT
NACA HS         EVKDIELVMSQ..ANVSRAKAVRALKNN....SNDIVNAIMEL
HYPK_HS         KKEDLELIMTE..MEISRAAAERSLREH....MGNVVEALIAL

NXF1 HS         QQEMLQAFSTQ..SGMNLEWSQKCLQDN....NWDYTRSAQAF
NXF2 HS         QQEMVQAFSAQ..SGMKLEWSQKCLQDN....EWNYTRAGQAF     TapCT
MEX67_SC        QLELLNKLHLE..TKLNAEYTFMLAEQS....NWNYEVAIKGF

EFTs_HS         LLMKLRRKTGY..SFVNCKKALETCGGD....LKQAEIWLHKE     EFTsNT
```

Figure 3-17 Alignment of representatives from the various UBA subfamilies. The names of the UBA subfamilies are shown at the right. Asterisks indicate residues of yeast Cue2 interacting with Ub as derived from pdb:1OTR. The 'sec.struc.' line indicates the position of the α-helices found in that structure. 'JPred' line renders a JPred secondary structure prediction for a whole-UBA-family alignment. Positions invariant or conservatively substituted in at least 40% of the sequences are shown on black and grey background, respectively. If more than one UBA domain is present in a sequence, they are numbered serially.

3.6.1.3 Additional UBA-related domains

Besides the CUE domain, several other short homology domains turned out to be distantly related to the UBA domain and probably assume the same structural fold (see Table 3-28). The biological meaning of these findings, especially, if these domains really bind to Ub or a related modifier, remains an open question in some cases.

AriNT (Ariadne N-terminus): This domain is found at the N-terminus of ARIH1, the human orthologue of Ariadne from *D. melanogaster* and some other members of the Parkin finger triad family. In addition, a second shared domain is detectable downstream of the triad, here called AriCT (Ariadne C-terminus). As the AriNT domain is found exclusively in some of the putative Parkin finger triad based ligases, a role in binding to Ub is probable.

TtrapNT (TNF- and TRAF associated protein N-terminus): In human, eight proteins with a TtrapNT domain are found, while in yeast there is only one: the Cdc53 interactor Dcn1. Among the human proteins are Cezanne 1 and Cezanne 2, two deubiquitylating proteases of the OTU class. Further TtrapNT domain proteins harbour additional domains typical of the UPS, like UBX or UIM domains. Obviously, most TtrapNT domain proteins have a probable or confirmed role in the UPS, which makes the TtrapNT domain a novel UBA derivative with Ub-binding properties.

NACα**CT** (nascent polypeptide associated protein complex α subunit C-terminus): Another UBA related domain was detected at the C-terminus of NACα. The human NACα protein and its yeast orthologue Egd2 are associated with cytoplasmic ribosomes and regulate their attachment to the ER membrane (Beatrix, 2000, Shi, 1995). Additional human proteins with this domain are KIAA0363 and HYPK, an interactor of huntingtin. None of the proteins listed above have thus far been described in processes that involve Ub. As the NACαCT-domain is even found across Ub-less archeae, a general role for this domain in Ub-binding appears unlikely.

TapCT: Another UBA-like domain, which is probably not involved in Ub binding is found at the C-terminus of NXF1/TAP and NXF2/TAPL2, which are known nuclear RNA export factors. This domain is also present in the yeast orthologue, the RNA export factor Mex67.

EFTsNT (elongation factor Ts N-terminal): The *E. coli* elongation factor Ts has been crystallized and the structural resemblance of its N-terminus to the general UBA fold is obvious (Kawashima, 1996, Yuan, 2004). This domain does not bind to Ub. Instead, a general role for the EFTsNT domain is to interact with another elongation factor, EF-Tu, which has no similarity to Ub. However, an evolutionary relationship between UBA and EFTsNT seems real.

Table 3-28 Proteins with UBA-related domains from yeast and humans. Orthology assignments are indicated by grey shading.

AriNTerm			
Yeast		**Human**	
Gene name	**ORF**	**Gene name**	**Uniprot number**
YKR017C	YKR017C	ARIH1	Q9Y4X5
		ARIH2	O95376
-	-	PARC	Q8IWT3
-	-	ANKIB1	Q6P3S9
TtrapNT			
Yeast		**Human**	
Gene name	**ORF**	**Gene name**	**Accession number**
DCN1	YLR128W	RP42/SCRO	Q96GG9
		C13orf17	Q6PH85
-	-	TTRAP/EAP2	O95551
-	-	ZA20D1/Cezanne1	Q6GQQ9
-	-	LOC124402	Q8TB05
-	-	C15orf16/Cezanne2	Q8TE49
-	-	MGC29814	Q8IYN6
-	-	NSFL1C	Q9UNZ2
-	-	NSFL1C-like	GENSCAN00000067370H
NACaCT			
Yeast		**Human**	
Gene name	**ORF**	**Gene name**	**Uniprot number**
EGD2	YHR193C	NACA	Q13765
-	-	KIAA0363	O15069
-	-	HYPK	Q8WUW8
-	-	FKSG17	Q9BZK3
-	-	MGC71999	Q9H009
TapCT			
Yeast		**Human**	
Gene name	**ORF**	**Gene name**	**Uniprot number**
MEX67	YPL169C	NXF1/TAP	Q9UBU9
		NXF2/TAPL2	Q9GZY0

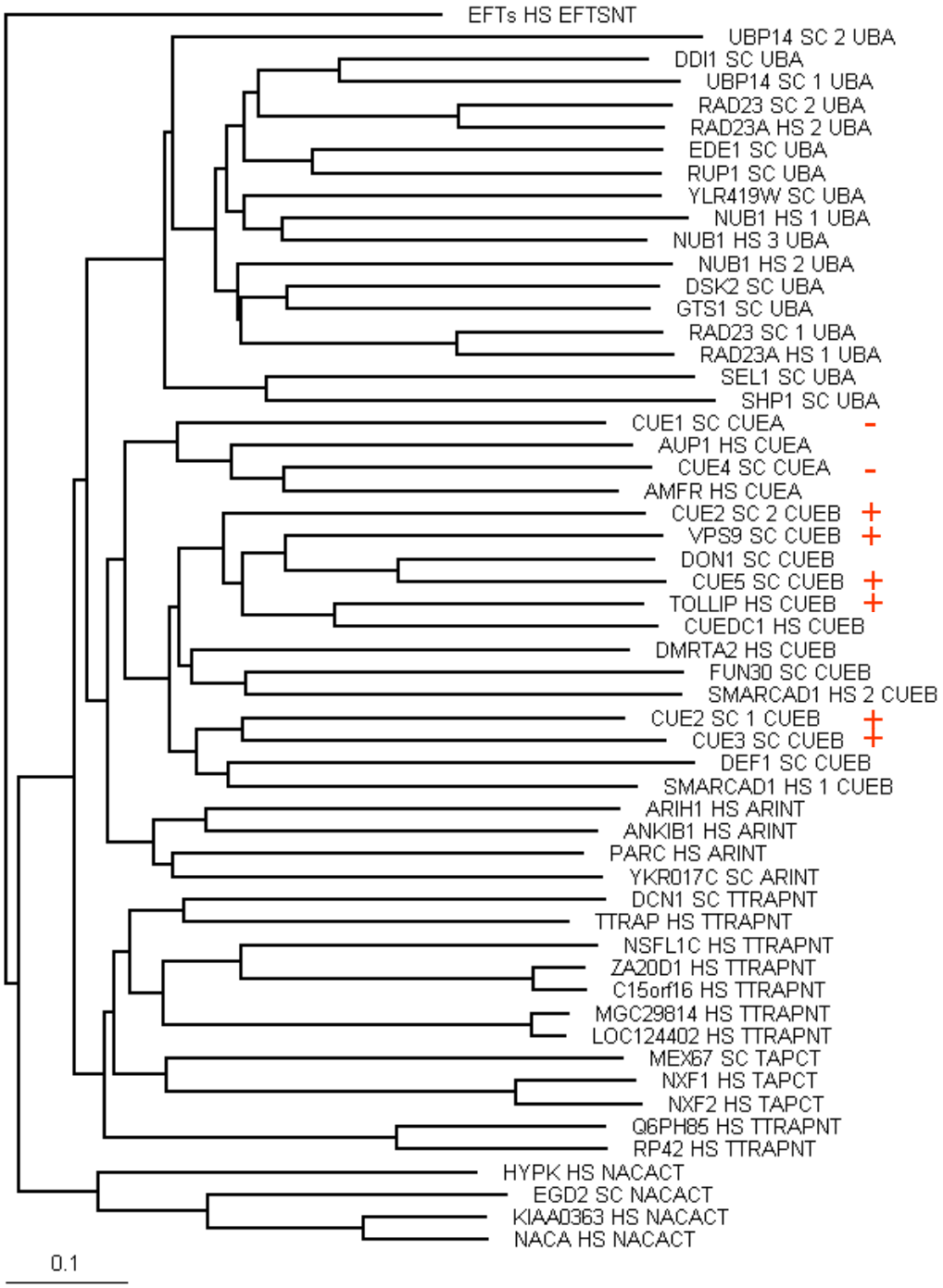


Figure 3-18 Subfamilies of the UBA superfamily. This dendrogram was generated from a multiple alignment covering the UBA and UBA-like domains of proteins depicted in Figure 3-17. Two-letter species abbreviation behind each protein name denote human (HS) or yeast (SC), respectively. In proteins with multiple domains, the domain number follows the species abbreviation. Red '+' and '-' symbols stand for domains positive or negative tested for interaction with Ub. The relationships within this tree do not necessarily reflect the exact evolutionary relationships between yeast and human proteins.

3.6.1.4 Human orthologues of yeast Snf1 lack the UBA domain

Snf1 is a protein serine/threonine kinase required for cellular response to altered nutritional conditions, like nitrogen starvation or alternate carbon sources (Sanz, 2000). It

contains a UBA domain C-terminally of the kinase domain. The human orthologues are PRKAA1/AMPK and PRKAA2/AMPK2 according to comparisons of the kinase domain and a short homologous stretch at the C-terminus of these proteins. Indeed, Snf1 and the proposed orthologues are functionally related, as they belong to a stress response system (Farras, 2001). However, the UBA domain is absent in the human proteins while present in numerous plant and fungal orthologues of Snf1, e.g. Snf1 from *S. pombe* (data not shown) (Hartmann-Petersen, 2003). Reverse profile searches from the UBA domain of fungal Snf1 orthologues confirmed its relationship to the classic UBA domain, but did not match the human Snf1 orthologues.

3.6.2 The Ub-interacting motif (UIM)

Like the UBA domain, the UIM is a well established Ub-recognition module and most UIM proteins are components of the UPS or other Ub-dependent pathways like endocytosis. Besides Ub recognition, a noteworthy feature of the UIM is its contribution to keep Ub-signals in a monoubiquitin state in endocytosis (Di Fiore et al., 2003). Of the eight UIM proteins found in yeast, at least four are functionally associated with receptor endocytosis/protein sorting (Hse1, Ent1, Ent2, Vps27) (see Table 3-29 for a complete list) (Bilodeau, 2003, Di Fiore et al., 2003). These proteins have different domain structures as depicted in Figure 3-19. They could be assigned to human orthologues that have the UIM conserved (see Table 3-29). Ub-dependent endocytosis of receptors and following protein sorting events seem to function similarly in yeast and human. However, not all of them appear to rely on the UIM as Ub-signal recognition motif, as Ede1, the yeast orthologue of the human endocytosis protein EPS15, has the UIM tandem replaced by a single UBA domain (see Figure 3-19).

Other proteins that have a UIM are the proteasomal subunit Rpn10 in yeast and its human orthologue PSMD4 (formerly S5a), as well as several DUBs, Ub-ligases and numerous so far uncharacterised proteins. The strong association of characterized UIM proteins with Ub-related processes suggests that the same might be true for other uncharacterised UIM-containing proteins. As an example, the functionally unexplored ataxin-3 protein, which depending on the splice variant has two or three UIM copies, was identified as a putative deubiquitylating enzyme similar to the USP type DUBs (see chapter 3.5.3) (Burnett et al., 2003, Scheel et al., 2003).

UIMs often occur in spatially adjacent tandems. As a single UIM is sufficient to interact with Ub, UIM tandems might allow linkage type recognition of Ub chains. Similar arrangements are observed in proteins with multiple UBA or CUE domains.

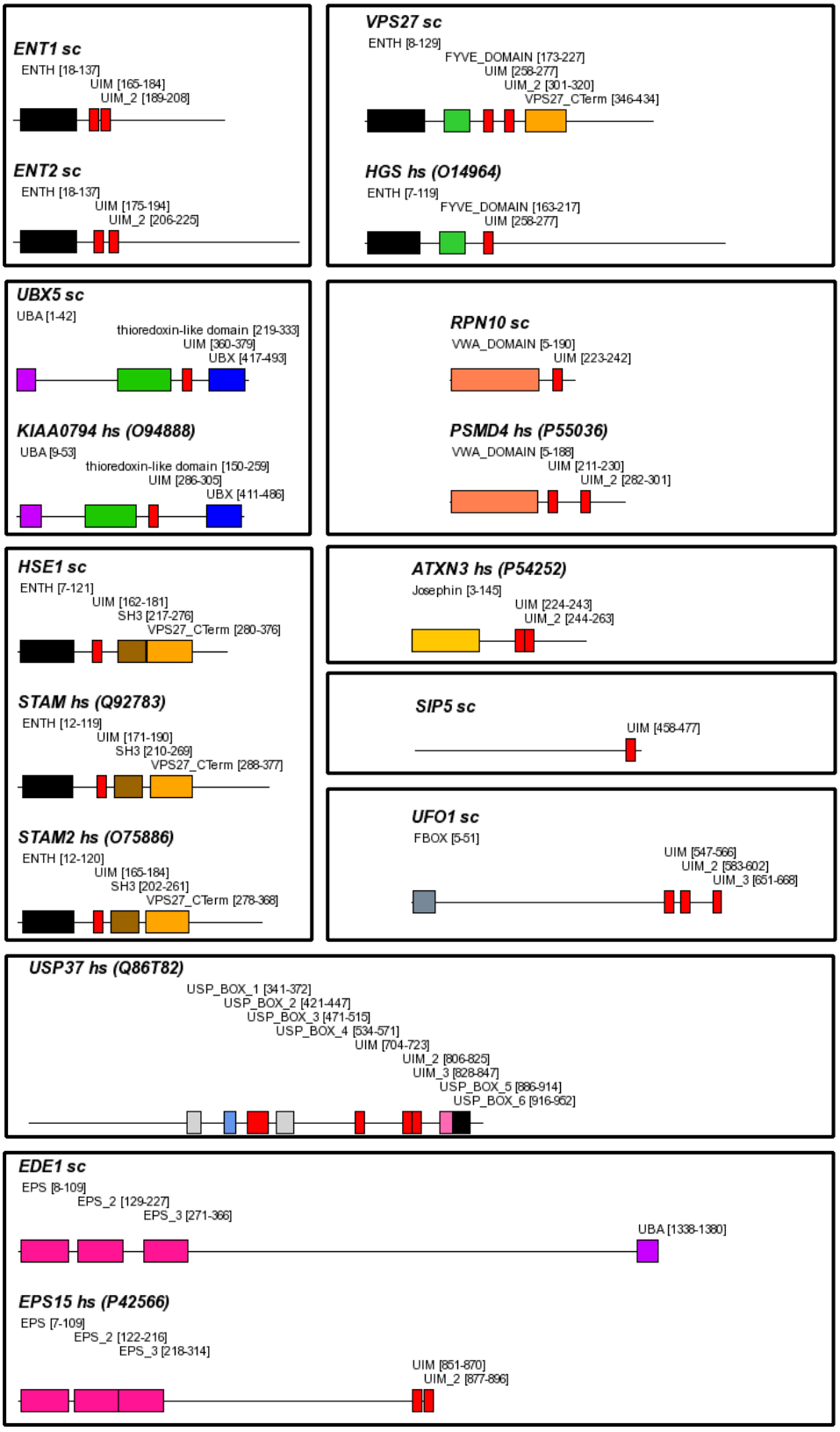


Figure 3-19 Domain topology of selected UIM proteins from yeast and human. UIMs are depicted as red boxes together with sequence coordinates. Other domains are coloured alternatively. Black frames indicate orthologues/paralogues. In the case of yeast Ent1/Ent2 human orthologues are not shown because of a highly similar domain topology. The human paralogue of EPS15, EPS15L1, lacks a UIM or analogous domain and was therefore omitted. The USP/UBP domain of human USP37 is shown as separate boxes as described in chapter 3.5.2.2 in order to make the inserted UIMs between USP/UBP box 4 and 5 visible.

Table 3-29 Proteins with UIM from yeast and human. Orthology assignments were made if possible.

UIM Yeast Gene name	ORF	Human Gene name	Uniprot number
ENT1	YDL161W	EPN1	Q9Y6I3
ENT2	YLR206W	EPN2	O95208
UBX5	YDR330W	EPN3	Q9H201
HSE1	YHL002W	HGS	O14964
RPN10	YHR200W	DNAJB2	P25686
UFO1	YML088W	EPS15	P42566
SIP5	YMR140W	EPS15L1	Q9UBC2
VPS27	YNR006W	ATXN3	P54252
Orthologues		ATXN3L	Q9H3M9
Yeast	**Human**	PSMD4	P55036
UBX5	KIAA0794	RAP80	Q5XKQ1
HSE1	STAM, STAM2	ANKIB1	Q6P3S9
RPN10	PSMD4	FLJ25555	Q6P5X6
VPS27	HGS	FLJ44474	Q6ZTN6
ENT1, ENT2	EPN1-3	UREB1	Q7Z6Z7
EDE1	EPS15, EPS15L1	ANKRD13	Q8IZ07
		LOC130617	Q8WV99
		STAM	Q92783
		STAM2	O75886
		USP28	Q96RU2
		USP25	Q9UHP3
		USP37	Q86T82
		KIAA0794	O94888

3.6.3 The UEV domain

The UEV domain, another proposed Ub-binding domain, is related to the E2 catalytic domain (Ubc domain). It has been dealt with only briefly in chapter 3.3 and will be described in more detail here. UEV proteins have been determined by profile searches with a profile based on the Ubc domain followed by selection of proteins without catalytic cysteine. The absence of the catalytic cysteine was carefully checked to rule out a mis-classification due to alignment errors.

Table 3-30 List of yeast and human UEV proteins and assignment of orthologues. The pseudogene is printed in italics.

Yeast UEV Gene name	ORF	Human UEV Gene name	Uniprot number
STP22	YCL008C	FTS	Q9H8T0
MMS2	YGL087C	TSG101	Q99816
Assignment of orthologues		UBE2V1	Q13404
Yeast	**Human**	UBE2V2	Q15819
STP22	TSG101	UEV3	Q8IX04
MMS2	UBE2V1, UBE2V2	*OTTHUMP00000030191*	*Q9NTT1*

The group of UEVs is much smaller than the group of Ubcs, comprising only Mms2 and Stp22/Vps23 in yeast and five proteins in human including TSG101, the orthologue of Stp22/Vps23 (see Table 3-30). Mms2 shares an orthology relationship to UBE2V1 and

UBE2V2, two paralogous human UEV proteins. Mms2 has a role in DNA damage repair, and the same can be assumed for its human orthologue. Yeast Stp22 and its human orthologue TSG101 are subunits of the ESCRT-1 complex acting on Ub-dependent sorting of proteins to multivesicular bodies (Garrus, 2001, Katzmann, 2001). Interestingly, TSG101 has been suggested to be a putative tumour suppressor (Li, 1997).

The crystal structure of Tsg101 has been solved by Pornillos et al. (Pornillos et al., 2002). According to the authors, structures of UEVs and active Ubc proteins are best superimposable in the region around the position occupied by the active cysteine in the active Ubcs. This finding is in agreement with the good conservation of the corresponding primary sequences.

3.6.4 The NZF domain

Two proteins, human Npl4 and yeast Vps36, have been demonstrated to bind to Ub via a particular type of C4 zinc finger domain (Meyer et al., 2002, Wang et al., 2003). Profile-based analysis of these zinc fingers clearly shows a relationship to the Ran-binding zinc finger in RanBP proteins (Yaseen, 1999). These proteins form a large family, most members of which do not appear to have a clear connection to the UPS. Therefore, it is currently questionable if all of these proteins are Ub-binding or if Npl4 and Vps36 are just special members of this family that are able to interact with Ub. The zinc fingers in Npl4 and Vps36 are not very similar to each other and do not constitute a clear subfamily with Ub-binding properties within the RanBP superfamily. See Table 3-31 for a list of NZF proteins. Remarkably, yeast Npl4 is devoid of this domain while the human orthologue has one. The opposite constellation is true for Vps36, where the human orthologue CGI-145 lacks the NZF domain.

Table 3-31 Proteins with an NZF-type zinc-finger from yeast and human. Orthology assignments were made if possible.

ZF_NZF					
Yeast		**Human**			
Gene name	**ORF**	**Gene name**	**Uniprot number**	**Gene name**	**Uniprot number**
NRP1	YDL167C	MDM4	O15151	YAF2	Q8IY57
VPS36	YLR417W	SOLH	O75808	RYBP	Q8N488
Orthologues		ZNF265	O95218	NEIL3	Q8TAT5
Yeast	**Human**	NUP153	P49790	NPL4	Q8TAT6
NPL4	NPL4	RANBP2	P49792	RNF31	Q96EP0
VPS36	CGI-145 (Q9Y3E3)	RBM6	P78332	UBCE7IP3	Q9BYM8
		RBM10	P98175	DKFZp434B1727	Q9H0E8
		MDM2	Q00987	MAP3K7IP2	Q9NYJ8
		TAB3	Q6VQR0	ZRANB1	Q9UGI0
		hRBCKL1-alpha	Q8IXF6		

3.6.5 The PAZ domain

The PAZ (poly-Ub associated zinc finger) domain was originally found in HDAC6, a histone deacetylase, where it is responsible for binding to poly-Ub (Hook, 2002, Seigneurin-Berny, 2001). Like the NZF domain, the PAZ domain belongs to the zinc-finger group of Ub-binding domains. So far, HDAC6 is the only member of the PAZ family with an established role in Ub-binding. However, as most of the remaining family members are DUBs belonging to the USP family, it is likely that the PAZ domain has a general role in the UPS. Due to its abundance in USPs, the PAZ domain is often also referred to as Zf-UBP, where 'UBP' is synonymous with 'USP'.

In human, 14 proteins with PAZ domain were detected, 12 of which are USPs (see Table 3-32). Only four PAZ proteins were found in yeast, which all have human orthologues. Remarkably, USP39 and Sad1 contain the PAZ domain. Both proteins are inactive USPs and Sad1 has a role in pre-mRNA splicing rather than in the UPS. The RING finger protein BRAP/RNF52 also contains a PAZ domain. It regulates MAP kinase activation by its Ras induced autoubiquitylation (Matheny, 2004). The Ub-binding PAZ domain in BRAP/RNF52 may facilitate the autoubiquitylation and similar events are probably true for the yeast orthologue Yhl010c, an uncharacterised ORF.

Table 3-32 Proteins with a PAZ domain from yeast and human. Orthology assignments were made if possible.

PAZ			
Yeast		**Human**	
Gene name	**ORF**	**Gene name**	**Uniprot number**
SAD1	YFR005C	USP3	Q9Y6I4
UBP14	YBR058C	USP5	P45974
UBP8	YMR223W	USP13	Q92995
YHL010C	YHL010C	USP16	Q9Y5T5
Orthologues		USP20	Q9Y2K6
Yeast	**Human**	USP22	Q9UPT9
SAD1	USP39	USP33	Q8TEY7
UBP14	USP5,USP13	USP39	Q9BV89
UBP8	USP22, USP27, USP51	USP44	Q9H0E7
YHL010C	BRAP/RNF52	USP45	Q70EL2
		USP49	Q70CQ1
		USP51	Q70EK9
		BRAP/RNF52	Q7Z569
		HDAC6	Q9UBN7

3.6.6 The GAT domain

3.6.6.1 Few GAT members in yeast and human

The GAT domain, which is present in GGA- and TOM1-homologues, is a homology domain that was initially thought of specifically interacting with Arf-type GTPases. However, the GAT domains of TOM1 and its close relatives are not able to interact with Arf (Katzmann, 2002). Shiba et al. have suggested to divide the initial GAT domain into two different subdomains with Arf and Ub-binding capacities, respectively (Shiba, 2003). While the Ub-binding subdomain is present in all GGA- and TOM1-homologues, the structural element essential for binding to Arf is only present in the GGA homologues.

There are only few yeast and human GAT domain proteins, which are either GGA- or TOM1-homologues (Table 3-33). In the GGA subfamily, there are no clear orthology relationships between yeast and human proteins. The two yeast proteins GGA1 and GGA2 form a group that is equally related to all three human GGA proteins, namely GGA1, GGA2 and GGA3. Yeast LSB5 is orthologous to all members of the human TOM1 subfamily.

Table 3-33 Proteins with GAT domain from yeast and human. Orthology assignments were made if possible.

GAT			
Yeast		Human	
Gene name	ORF	Gene name	Uniprot number
GGA1	YDR358W	TOM1	O60784
GGA2	YHR108W	TOM1L1	O75674
LSB5	YCL034W	TOM1L2	Q8TDE7
Orthologues		GGA1	Q9UJY5
Yeast	Human	GGA2	Q9UJY4
GGA1, GGA2	GGA1, GGA2, GGA3	GGA3	Q9NZ52
LSB5	TOM1, TOM1L1, TOM1L2		

3.6.6.2 The GAT domain and the UIM appear exchangeable

Interestingly, most GAT domain proteins contain a VHS domain (originally found in Vps27, Hrs, STAM) at their N-terminus. However, Vps27, Hrs and STAM themselves lack the GAT domain and contain a UIM at a position, where a GAT domain would be expected. Obviously, VHS domain proteins employ different Ub-binding domains or motifs for proper function. This observation is only one of many non-orthologous domain replacements within the UPS. Although the UIM and the GAT domain bind to Ub, it remains elusive if they are really equivalent or if they have different affinities to Ub.

3.6.7 Identification of a SUMO interaction motif

So far, several domains have been described that are well established recognition motifs for Ub. Sumoylation of proteins has been shown to be important for various cellular processes (Dohmen, 2004). By contrast, no generally valid SUMO interaction domain or motif is known so far. The mechanism by which sumoylation regulates these processes is still unknown, but physical interactions between SUMO moieties and some sort of recognition motifs are expected. Here, based on physical interactions between proteins and SUMO/Smt3 in yeast, a bioinformatical approach to identify potential SUMO interaction motifs was tested.

3.6.7.1 Working hypothesis

Given a set of SUMO interacting proteins, multiple binding modes may exist like for Ub and its recognizing motifs. A worst case scenario would be that each of the interactors employs a different binding mode. Different binding modes are usually equivalent to different binding motifs, so that no common sequence features might be present. The SUMO interactors may share a common sequence feature (a SIM, for SUMO interacting motif), which allows them to bind to SUMO. Such sequence features might be manifested as conserved sequence motifs or even as homology domains like the UBA domain. If so, it should be possible to identify a SIM in SUMO interacting proteins by sequence comparison methods. Problems would be a high divergence of a common SUMO binding motif or large inserts of primary sequence between the residues mediating the interaction with SUMO. Both would make a SIM nearly invisible. To circumvent these problems, the profile method was chosen due to its high sensitivity and ability to cope with large insertion between conserved sequence blocks as seen in chapter 3.5.2.3.

3.6.7.2 Working scheme for SIM identification

A schematic overview of all steps is depicted in Figure 3-20. First, interaction data for yeast SUMO had to be compiled. A study of K. Uzunova and J. Dohmen served as a primary source for SUMO interacting partners, which were isolated in a yeast-two-hybrid screen (K. Uzunova and J. Dohmen, unpublished results). An advantage for the following analysis was that in these SUMO interactors the regions essential for SUMO binding had already been narrowed to some extent. Additional SUMO interactors were extracted from a compiled set of yeast protein-protein interactions (Ito et al., 2001, Uetz et al., 2000).

In the second step, orthologues for each of the SUMO interactors were retrieved from various databases including numerous fungal sequence databases published recently. At this stage, BLAST searches were sensitive enough to find a sufficient number of orthologues, from

which alignments were constructed afterwards. According to the working hypothesis, a putative SIM should now be located somewhere within such an alignment if conserved across the species.

In the next step, profiles based on these alignments were constructed in order to screen the remaining SUMO interaction partners for the yet hidden SIM. Only those regions were used that had been shown to be sufficient for SUMO binding according to the results from the screen of K. Uzunova and J. Dohmen carried out with truncated proteins. Positions marking transitions from conserved to non-conserved regions were chosen as boundaries in the dissection of each alignment. This procedure resulted in multiple profiles per alignment and therefore per SUMO interactor family. The profiles were then used for searches in a sequence database including the sequences of the SUMO interactors.

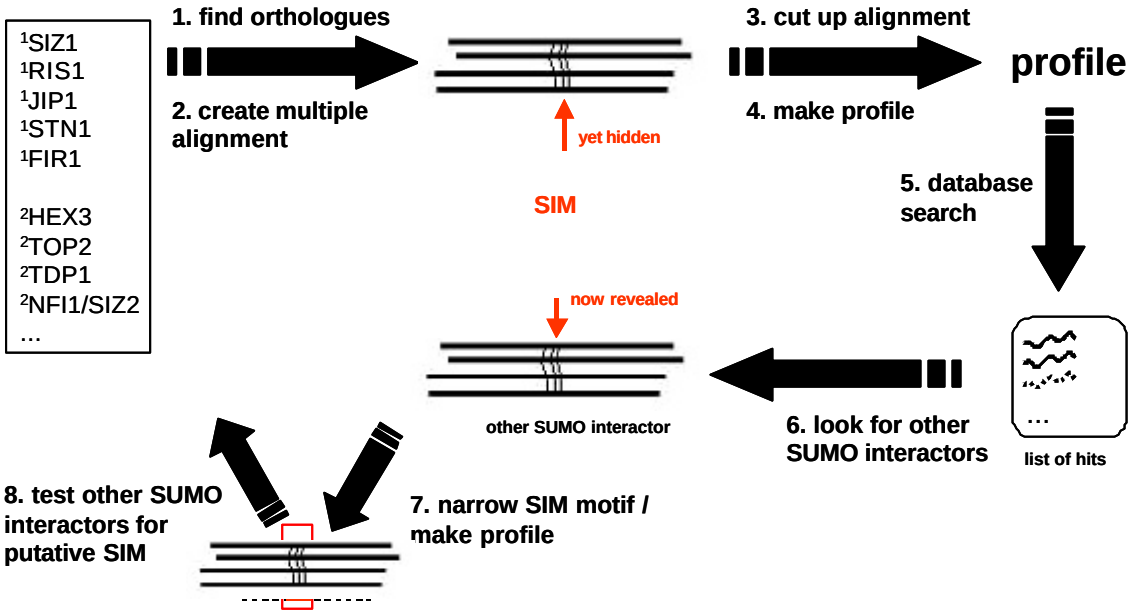


Figure 3-20 Working scheme for SIM identification. '1' SUMO interactors provided by K. Uzunova and J. Dohmen. '2' SUMO interactors found in the literature.

3.6.7.3 Integration of multiple criteria to evaluate putative SIMs

If any of the interactors or one of their orthologues was detected during such a search, the corresponding region of putative homology had to be validated by different aspects. In several instances, the profiles succeeded in finding one of the other interactors. However, the profile scores alone turned out to be inappropriate for evaluating the matches due to the extreme shortness of the matched regions. Instead, multiple criteria were established that had to be met by a matching sequence. First, only matches lying in SUMO binding regions were regarded. Another criterion took into account the phylogenetic distribution of the putative SIM across the orthologues of an interactor. For that purpose, the SIM was tested for conservation in each

orthologue and compared against an 18S rRNA derived species tree (construction see chapter 2). This criterion helped to uncover random hits and, at the same time, allowed to rank putative SIMs according to their phylogenetic distribution. A third major aspect in evaluating a match was the conservation of the sequence up- and downstream of the matching region. For that purpose, it was tested if the matching region was just part of a larger domain or if it was embedded as a 'conserved island' in an otherwise unstructured sequence neighbourhood. The latter constellation may indicate some kind of structural and functional autonomy and was therefore of particular interest. In contrast, matching regions that were obviously part of a larger homology domain might represent buried stretches with a structural role and are probably unable to interact with SUMO.

3.6.7.4 Identification of a putative SUMO interaction motif

In one of the database searches with a profile based on an MSA covering residues 248-407 of yeast Ris1, human Uba2 and yeast Fir1 were detected. Both proteins were of particular interest, as the region matched in Fir1 is part of a fragment that has been shown to bind SUMO in the screen of K. Uzunova and J. Dohmen. Likewise, human Uba2 is a subunit of the heterodimeric SUMO-E1 and Uba2 has been shown to interact with SUMO prior to linking it covalently (Gong, 1999). Interestingly, Fir1 and Uba2 matched to the same ~10 residue long region in the Ris1-derived profile. Subsequently, corresponding regions from various orthologues of Ris1, Fir1 and Uba2 were aligned in order to define a shared motif that might be SUMO binding (see Figure 3-20).

The extracted motif spans only 10 residues and is bipartite concerning the physicochemical properties of the residues (see Figure 3-22). At the N-terminus, a usually four amino acids long hydrophobic patch is present, which is C-terminally flanked by a stretch of acidic residues often containing serine residues. Within the Ris1 orthologues, a serine seems to separate the distinct halves, but this serine is absent in Uba2 and Fir1 proteins. The phylogenetic distribution clearly shows that this motif is conserved across the Fir1, Ris1 and Uba2 orthologues. This motif could even be traced back to human Uba2 (see Figure 3-21). In the following steps, distinct variants and combinations of the motif were converted to profiles in order to test the remaining known SUMO interactors and especially their mammalian orthologues for the presence of this motif.

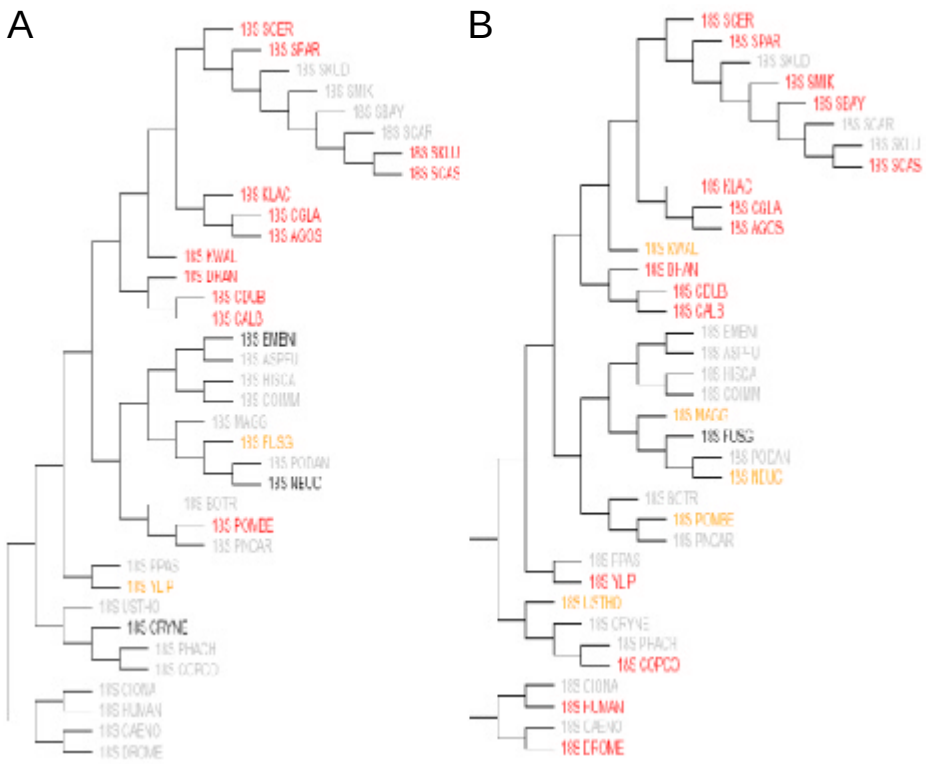


Figure 3-21 Phylogenetic distribution of a given SIM across orthologues of Ris1 (A) and Uba2 (B). Each tree renders the relative evolutionary relationship of selected species based upon their 18S rRNA. The phylogenetic distribution was mapped on this tree by usage of a colour code. Red labels indicate the presence of a SIM in the orthologous gene of the corresponding species. Other colour codes are yellow = ‚SIM present, but degenerate' , black' = absent, ‚grey' = not tested.

3.6.7.5 The putative SIM is present in numerous SUMO interactors

SIM-based profiles successfully detected a SIM-like stretch in the fragment of Siz1 sufficient for binding to SUMO and in the Siz1 homologue, Siz2 (see Figure 3-22). Both proteins are known SUMO ligases (Johnson et al., 2001, Takahashi et al., 2003), which makes the presence of a SUMO binding motif plausible. In Nis1, only a highly divergent variant of the original SIM could be found, while Stn1 has no SIM as defined here.

For Nis1, the putative SIM is located in the short ~60 residue C-terminal fragment, which is still able to interact with SUMO, and is conserved across the close relatives of *S. cerevisiae*, the Saccharomyces *sensu strictu* branch. Other interesting matches are listed in Figure 3-23 together with the SIM coordinates. Among these matches was Wss1 (weak suppressor of Smt3), which is involved in protein sumoylation and might play a role in DNA damage response (Biggins, 2001) (O'Neill, 2004). However, the SIM in Wss1 seems to be restricted to fungal species. Noticeably, Ubc9, which binds to and conjugates SUMO (Johnson et al., 1997), has a stretch quite similar to the SIM directly adjacent to its catalytic cysteine (data not shown). An example for a metazoan-specific SUMO-interactor exhibiting the SIM is human

Daxx (death-domain-associated protein 6) (Ryu, 2000). Daxx has two putative SIMs at its extreme N- and C-terminus, respectively, with an sumoylation site in between (data not shown).

```
Ris1_YEAST       370 SIIILSDEDE
Ris1_KLAC         66 QVIVVSSEEE
Ris1_POMBE       185 AVIVVSDSES
Ris1_CALB        123 EVIALSDSDD

Fir1_YEAST       758 EVILLDEDED
Fir1_KLAC        753 EVIVLDPPED

Uba2_2_YEAST     582 GIVILDDDEG
Uba2 2 KLAC      576 ETLIVDDEPA
Uba2_2_HUMAN     586 DVLIVDSDEE

Uba2 1 YEAST     508 SIILFSDEEGD
Uba2_1_KLAC      501 TFLLYKDEEME

Nis1/Jip1_YEAST  390 PIIIPDSQDD
Nis1/Jip1_KLAC   398 PIVISDNEDA

Siz1_YEAST       480 PIIINLDSDDDE
Nfi1/Siz2 YEAST  470 PEIISLDSSDDE
PIAS1_MOUSE      459 VIDLTIDSSSDE
PIAS3 HUMAN      447 VIDLTIESSSDE
PIAS4_HUMAN      465 VVDLTLDSSSSS
```

Figure 3-22 Potential SUMO interaction sites in known SUMO interactors. Hydrophobic residues are printed in green, negatively charged ones in blue.

```
start   end      ORF        ... start   end      ORF
 758 -  767 yp|YER032W [FIR1]    16 -   25 yp|YDL235C [YPD1]
  23 -   32 yp|YDL013W [HEX3]   822 -  831 yp|YBL052C [SAS3]
 480 -  492 yp|YDR409W [SIZ1]    54 -   63 yp|YMR131C [RRB1]
 471 -  482 yp|YOR156C [NFI1]   454 -  463 yp|YMR277W [FCP1]
 370 -  379 yp|YOR191W [RIS1]   306 -  315 yp|YOL054W [PSH1]
 390 -  401 yp|YNL078W [NIS1]    81 -   90 yp|YKR062W [TFA2]
 246 -  255 yp|YHR134W [WSS1]  1939 - 1948 yp|YDR457W [TOM1]
 582 -  591 yp|YDR390C [UBA2]   270 -  279 yp|YMR037C [MSN2]
 453 -  462 yp|YOR123C [LEO1]   140 -  149 yp|YBR049C [REB1]
 217 -  225 yp|YDR054C [CDC34]  249 -  258 yp|YIL131C [FKH1]
 555 -  564 yp|YER049W YER049W  883 -  892 yp|YCR057C [PWP2]
  62 -   71 yp|YDL153C [SAS10]  632 -  641 yp|YMR224C [MRE11]
 521 -  530 yp|YBR215W [HPC2]   614 -  623 yp|YAL043C [PTA1]
 180 -  188 yp|YDR330W [UBX5]
```

Figure 3-23 Extended list of yeast proteins with a potential SIM. Numbers denote the SIM coordinates.

3.6.7.6 The SIM is orthologous to a previously defined SUMO binding motif in PIAS2

Analysis of the phylogenetic distribution of the SIM in Siz1 revealed its conservation across fungal and metazoan species. This is of particular interest, as one of the human orthologues of Siz1, PIAS2 (also termed PIASx), has been described to carry a SUMO binding site (Minty, 2000). Therefore, it had to be examined if the SIM defined here was different from the one reported by Minty et al.. According to this publication, the SUMO binding site in PIAS2/PIASx and further proteins depends on a serine-x-serine triplet flanked by hydrophobic residues upstream and acidic ones downstream. The importance of both serine residues could not be confirmed here, as the serines are not conserved in most of the yeast SUMO interactors

and their orthologues. In some cases, serine residues were found at positions normally occupied by aspartate or glutamate. It is conceivable that a phosphorylation of these serines allows modulation of the SUMO binding ability, as phosphoserine might mimic the acidic amino acids in the acidic patch.

Yeast-two-hybrid experiments with the motif isolated from PM-Scl75 have demonstrated the ability of this motif to bind to SUMO (Minty et al., 2000). Various mutagenesis experiments carried out in the same work suggested that the hydrophobic patch, the acidic patch as well as the spacing between the two patches are crucial for SUMO binding. These findings fully underline the SIM and its boundaries as defined in this work.

In a publication of Song et al., a SUMO binding motif related to the SIM described here has been reported after the analysis performed in this work (Song, 2004). Song et al. could define the residues of the motif involved in SUMO interaction in more detail by NMR spectroscopic means As a result, the residues of the hydrophobic patch contribute mainly to the SIM-SUMO interaction, while the remainder of the motif is less important. These results are in good agreement with the conservation pattern of the SIM, whose most conserved part is the hydrophobic patch, while the acidic stretch often has polar residues like serine interspersed.

3.7 Proteasome

3.7.1 The 20S proteasome

Both yeast and human have an identical 20S proteasome consisting of 14 α-subunits and 14 β-subunits, which in turn are encoded by 7 α-subunit genes and 7 β-subunit genes in yeast. In human, three β-subunit genes and one α-subunit gene have been duplicated during evolution and at least the three β-subunits are found exclusively in the immunoproteasome (Table 3-34).

3.7.1.1 The immunosubunits are paralogues of the catalytic 20S proteasome subunits

Mammalian genomes encode three additional immunosubunits not found in yeast (Gille et al., 2003). These subunits, termed PSMB8/β5i, PSMB9/β1i and PSMB10/β2i occupy the positions of PSMB5/β5, PSMB6/β1 and PSMB7/β2 of newly synthesized 20S immunoproteasomes. According to sequence comparisons, PSMB5/PSMB8, PSMB6/PSMB9 as well as PSMB7/PSMB10 form closely paralogous pairs (see Table 3-34). At least for *D. melanogaster* duplications of core proteasomal subunits are also known (Ma, 2002).

3.7.1.2 Human PSMA7/α4 has recently been duplicated

While screening the human sequence database for proteasomal subunits, a remarkable finding was the existence of a paralogous sequence of the subunit PSMA7/α4 encoded on chromosome 18. This sequence is already present in the UniProt database, where it is referred to as PSMA7L/PSMA8. It is located on chromosome 20 and expressed in testis, so it might be a subunit of a putative testis-specific proteasome variant. The testis-specificity of PSMA7L/PSMA8 is an interesting parallel to the duplicated proteasomal core subunits of *D. melanogaster* mentioned above, of which several are specifically expressed in the male germline (Ma et al., 2002). Orthologues of PSMA7L/PSMA8 were also detected in mouse (Q9CWH6), zebrafish (Q6P0I2) and goldfish (Q9PTW9), indicating that this proteasomal subunit might play a role in vertebrates in general. Orthologues in insects or nematodes could not be found, which suggests that the duplication of the PSMA7 gene was a recent event in evolution.

Table 3-34 Components of the 26S proteasome. This table contains all yeast and human proteasome subunits according to literature. Subunits are sorted according to their membership of the proteolytic core, the base or the lid. The β-subunits are subdivided into catalytic and non-catalytic ones. Recently discovered novel subunits are listed as well and references to each of these proteins will be given in the text. Orthologues are found within a table row. For PRE6, PRE2, PRE3 and PUP1, respectively, two human proteins were identified as orthologues.

Yeast		Human		
20S proteasome core particle, a subunits				
Gene name	**ORF**	**Gene name**	**Uniprot number**	
PRE5	YMR314W	PSMA1/α6	P25786	
PRE8	YML092C	PSMA2/α2	P25787	
PRE10	YOR362C	PSMA3/α7	P25788	
PRE9	YGR135W	PSMA4/α3	P25789	
PUP2	YGR253C	PSMA5/α5	P28066	
SCL1	YGL011C	PSMA6/α1	P60900	
PRE6	YOL038W	PSMA7/α4	O14818	
		PSMA8/PSMA7L	Q8TAA3	
20S proteasome core particle, non-catalytic b subunits				
Gene name	**ORF**	**Gene name**	**Uniprot number**	
PRE7	YBL041W	PSMB1/β6	P20618	
PRE1	YER012W	PSMB2/β4	P49721	
PUP3	YER094C	PSMB3/β3	P49720	
PRE4	YFR050C	PSMB4/β7	P28070	
20S proteasome core particle, catalytic b subunits				
Gene name	**ORF**	**Gene name**	**Uniprot number**	
PRE2	YPR103W	PSMB5/β5	P28074	
		PSMB8/LMP7/β5i	P28062	
PRE3	YJL001W	PSMB6/β1	P28072	
		PSMB9/LMP12/β1i	P28065	
PUP1	YOR157C	PSMB7/β2	Q99436	
		PSMB10/MECL1/β2i	P40306	
19S regulatory particle, base subunits				
Gene name	**ORF**	**Gene name**	**Uniprot number**	
RPT1	YKL145W	PSMC2/S7	P35998	AAA
RPT2	YDL007W	PSMC1/S4	P62191	AAA
RPT5	YOR117W	PSMC3/S6a	P17980	AAA
RPT3	YDR394W	PSMC4/S6b	P43686	AAA
RPT6	YGL048C	PSMC5/S8	P62195	AAA
RPT4	YOR259C	PSMC6/S10b	P62333	AAA
RPN1	YHR027C	PSMD2/S2	Q13200	PC-REP
RPN2	YIL075C	PSMD1/S1	Q99460	PC-REP
RPN10	YHR200W	PSMD4/S5a	P55036	UIM
19S regulatory particle, lid subunits				
Gene name	**ORF**	**Gene name**	**Uniprot number**	
RPN3	YER021W	PSMD3/S3	O43242	PCI
RPN5	YDL147W	PSMD12/p55	O00232	PCI
RPN6	YDL097C	PSMD11/S9	O00231	PCI
RPN7	YPR108W	PSMD6/S10	Q15008	PCI
RPN9	YDR427W	PSMD13/S11	Q9UNM6	PCI
RPN12	YFR052W	PSMD8/S14	P48556	PCI
RPN11	YFR004W	PSMD14	O00487	MPN+
RPN8	YOR261C	PSMD7	P51665	MPN
other putative subunits of the proteasome				
Gene name	**ORF**	**Gene name**	**Uniprot number**	
UBP6	YFR010W	USP14	P54578	USP
HUL5	YGL141W	UBE3B	Q9BXZ4	HECT
ECM29	YHL030W	KIAA0368	O15074	HEAT repeats
RPN13	YLR421C	ADRM1	Q16186	
SEM1	YDR363W-A	SHFM1/DSS1	P60896	
NAS2	YIL007C	PSMD9	O00233	PDZ
	-	PSMD10	O75832	Ankyrin repeats
-	-	PSMD5/S5b	Q16401	HEAT repeats

3.7.2 Subunits of the 19S regulatory particle

The 19S regulatory particle consists of a hexameric ring of six AAA ATPases plus three non-ATPase subunits and a lid complex built up from six subunits harbouring a so-called PCI domain plus two subunits containing MPN domains ('6+2' stoichiometry). For each of the nine base and eight lid subunits clear 1:1 orthology assignments between yeast and human could be determined (Table 3-34). The PCI subunits have some interesting properties, because they form the scaffold of the so-called PCI complexes in general and moreover, mediate physical interactions with a variety of additional proteins, such as kinases, deubiquitylating enzymes, RING/cullin based E3s or other PCI complexes. Therefore, a more detailed sequence analysis of the PCI subunits has been performed as described below.

3.7.3 Comprehensive analysis of the PCI subunits of the proteasomal lid, the CSN and the eIF3

Note: the following work has been published as "Prediction of a common structural scaffold for proteasome lid, COP9-signalosome and eIF3 complexes." by Scheel, H. and K. Hofmann in BMC Bioinformatics 6(1): 71 (Scheel, 2005).

3.7.3.1 Determining subunits of the PCI complexes

The proteasomal lid subcomplex, the COP9 signalosome (CSN) and the eIF3 complex share the property that all of them have a common scaffold made of PCI and MPN proteins. The PCI proteins are one of the main building blocks of the three PCI-based complexes, a fact already suggested by their high portion. There are several hints that the PCI subunits are crucial for proper complex assembly (Freilich, 1999, Lier, 2002, Tsuge, 2001, Valasek, 2001). The MPN subunits of the three complexes are rather well conserved and the detection of MPN domains and their boundaries was relatively straightforward. By contrast, the degree of conservation between PCI subunits is highly variable. Sequence similarity between the corresponding subunits of proteasome lid and CSN was generally easy to spot, while the detection of similarity between other paralogous PCI subunits typically required the generalized profile method. A particular challenge was the detection of the highly divergent PCI domains in the yeast CSN-like complex (Maytal-Kivity, 2003) and those of the eIF3 complex, where only three PCI subunits could be detected in the initial survey (Hofmann et al., 1998). Due to these difficulties, it is to be expected that there are still a number of highly divergent PCI domain proteins in eukaryotic genomes, which have eluded detection so far. A second issue in the bioinformatical definition of the PCI domain concerns the position of its N-terminal boundary. In general, homology domains are thought to correspond to structural domains in the sense of

autonomous folding units; they are typically characterized by a pronounced loss of sequence similarity at the domain boundaries. While this is true for the PCI domain C-terminus, the N-terminal domain boundary is blurred through a gradual decay in sequence similarity instead of a sharp drop. As a consequence, different PCI domain boundaries have been used in the literature (Aravind et al., 1998, Hofmann et al., 1998) and in various domain databases like PROSITE (Hulo et al., 2004), Pfam (Bateman et al., 2004) and SMART (Letunic, 2004). The corresponding accession numbers are PS50250, PF01399 and SM00088, respectively.

During an exhaustive bioinformatical analysis of PCI proteins, two independent results were obtained jointly suggesting that a structure-based redefinition of the PCI domain is appropriate: on one hand, multiple instances of TPR-like repeats were detected in the N-terminus of many PCI proteins, which suggests that the homology between the proteasome and CSN components is not restricted to the PCI domain itself. On the other hand, a previously overlooked PCI domain was revealed in the novel eIF3 subunit eIF3k (Mayeur, 2003). Most interestingly, an X-ray structure of eIF3k has been published recently (Wei, 2004). Based on this structure and on alignment data, a bipartite consensus model is suggested for the canonical PCI proteins, consisting of a C-terminal 'winged helix' domain preceded by an extended helical repeat region. This model has been used to re-evaluate some bioinformatical and experimental findings that have been enigmatic so far.

3.7.3.2 TPR-like helical repeats in PCI proteins

In most PCI proteins, the canonical PCI domain occupies a region of approximately 190 residues close to the carboxy-terminus of the sequence. The N-terminal non-PCI portion of the proteins is moderately conserved between species and only poorly conserved between different PCI subunits - even between the analogous subunits of the lid and the CSN. Upon submitting those PCI proteins to profile- or HMM-based domain detection services, no significant matches were obtained for the N-termini of the proteins. However, the PROSITE profile for the tetratrico-peptide repeat (TPR) yielded a number of closely sub-significant matches in multiple PCI proteins, e.g. Rpn7 from *S. bayanus* (P value = 0.01, Ref (Kellis et al., 2003)) and Csn1 from *E. histolytica* (P value = 0.06, Uniprot: Q8WQ58). The TPR repeat family (D'Andrea, 2003) is very heterogeneous, and TPR motif descriptors such as the PROSITE profile are known to miss several instances of *bona fide* TPR repeats. Upon closer inspection, most PCI proteins exhibit multiple regions of similarity to profiles derived from established TPR repeats (matches schematically shown in Figure 3-24), although the similarity scores for each of the

single regions do not reach statistical significance. No relevant similarity scores were obtained for other helical repeat motifs, such as HEAT or Armadillo repeats.

To further investigate if a TPR-like structure should be assumed for the N-terminal portions of all PCI proteins, a secondary structure prediction was performed for each of the protein families individually. To that aim, multiple alignments were constructed for representative members of each subunit family and submitted to PHD and JPred prediction servers (Cuff et al., 1998, Rost et al., 2003). As a result, all PCI subunits of lid and CSN are predicted to adopt an all-helical secondary structure upstream of the PCI domain. Interestingly, these helical regions merge seamlessly into the PCI domain, at least if the longer PCI versions of PROSITE and Pfam are used. This finding is in agreement with the observation of several regions with weak TPR-similarity within the N-terminal part of the PCI domain itself (see Figure 3-24). Further support for a TPR-like structure comes from a sequence-based fold recognition for lid and CSN subunits using the Superfamily-service (see Table 3-35) (Gough, 2001). Several subunits like Rpn7 from yeast and human Csn1 were found to have good scores for the TPR fold upstream of the PCI domain.

The predicted all-helical secondary structure of the non-PCI portion of lid and CSN subunits consists of several short helices that appear to occur in pairs. To test whether those bi-helical segments correspond to the structural elements of a TPR-like repeat, several examples were selected starting immediately upstream of the predicted PCI domains. When multiple alignments of those bi-helical segments were used for profile construction and in subsequent database searches several *bona fide* TPR proteins were found to match within the TPR region. With the bi-helices being in the correct TPR register, these segments were also classified as TPR-like. No matches to established HEAT- or Armadillo-repeat proteins were found, demonstrating that the scores are not just caused by an arbitrary helical repeat arrangement.

It should be pointed out that none of the singular observations described above is able to prove a statistically significant sequence relationship between the N-terminal portions of PCI proteins and true TPR-repeats. Taken as a whole, the results strongly suggest that there is a general tendency of PCI domains to be preceded by an α-helical repeat structure that has at least some specific relationship to the tetratrico-peptide repeat.

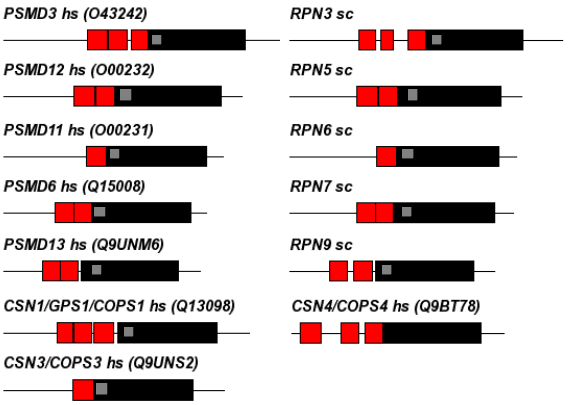


Figure 3-24 TPR-like motifs upstream and inside the PCI domain. Besides the common PCI domain (black), short stretches of ~35 aa each are depicted in red. These stretches show weak to medium similarity to TPR segments in established TPR proteins and merge seamlessly into the PCI domain in several PCI subunits. A grey box indicates a TPR-like stretch within the PCI domain.

Table 3-35 Fold prediction with the 'Superfamily' webserver for pre-PCI regions. The Superfamily webserver (Gough et al., 2001) proposes a TPR-like fold for PCI upstream regions in many PCI proteins.

subunit	species	position	E-value	Superfamily	structure	comment
Rpn7	yeast	131-200	3.2e-06	TPR-like	1hz4	
Csn1	human	65-218	1.1e-04	TPR-like	1hz4	below cut-off
Csn4	human	80-208	8.2e-03	TPR-like	1qqe	below cut-off
Rpn3	human	51-165	3.3e-02	TPR-like	1qqe	below cut-off
Rpn9	human	59-150	1.3e+00	TPR-like	1qqe	below cut-off

3.7.3.3 A previously unrecognised PCI domain in eIF3k

In the first surveys of recognizable PCI domains, only three PCI subunits of the eIF3 complex had been detected (Hofmann et al., 1998). More recently, a number of novel eIF3 components have been identified: eIF3j (Valasek, 2001), eIF3k (Mayeur et al., 2003) and eIF3l (Morris-Desbois, 2001). Among these novel subunits, only eIF3l has been reported to harbour a PCI domain (Morris-Desbois et al., 2001), interestingly also preceded by a TPR-region. In order to find further indications of divergent PCI domains, a thorough profile analysis of all uncharacterised eIF3 subunits was performed.

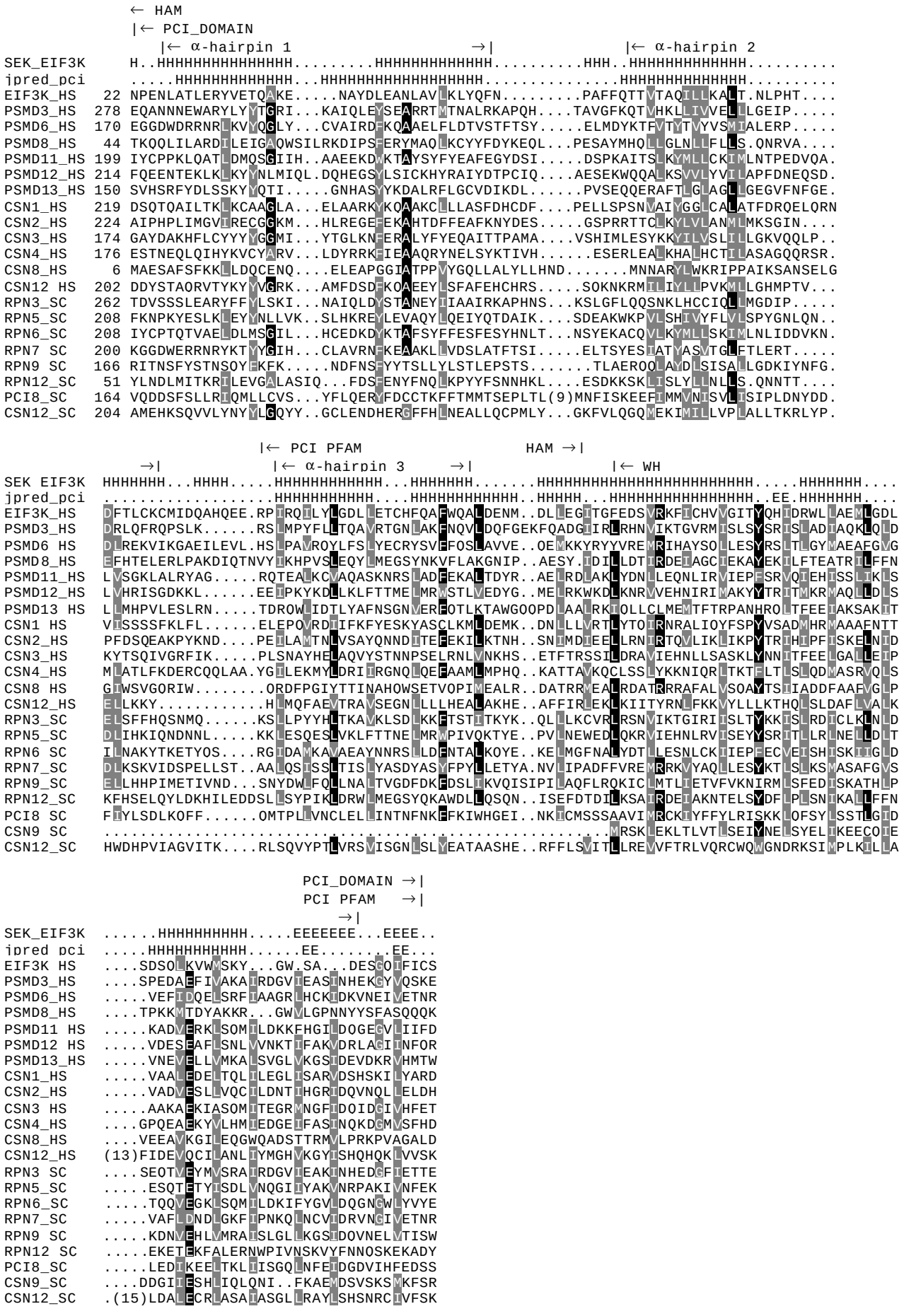
Table 3-36 PCI complexes and their subunit correspondence.

Domain	human lid	yeast lid	human CSN	yeast CSN	human eIF3
PCI	PSMD6	Rpn7	Csn1	Pci8/Csn11	eIF3(a,c,e,l)
PCI	PSMD11	Rpn6	Csn2	Rri2/Csn10	
PCI	PSMD3	Rpn3	Csn3	-	
PCI	PSMD12	Rpn5	Csn4	-	
PCI	PSMD13	Rpn9	Csn7a/Csn7b	Csn9	
PCI	PSMD8	Rpn12	Csn8	-	eIF3k
PCI	-	-	Csn12*	Csn12*	-

A generalized profile was constructed from the conserved portion of representative eIF3k orthologues from vertebrates, invertebrates, plants and fungi. After a scaling step, the

resulting profile was run against a non-redundant protein database. Apart from the eIF3k proteins already used for profile construction, the only other sequences matching with significance were selected PCI subunits of the proteasome and the CSN, among them rice Csn8 (p=0.01) and the Drosophila Rpn12 homologue (p=0.05). All of the twenty top-scoring sequences could be identified as either Csn8- or Rpn12-homologues. As shown in Table 3-36, Csn8 and Rpn12 are the corresponding PCI subunits in the CSN and the lid, respectively. Csn8 and Rpn12 are the most divergent PCI subunits of the proteasome and the signalosome, respectively, and their PCI domains appear to be shorter than that of the more typical family members. These observations provide good bioinformatical evidence that eIF3k is the fifth PCI-containing subunit of the eIF3 complex and most likely a direct analogue of Csn8 and Rpn12 (Figure 3-25 and Table 3-36).

Figure 3-25 on the following page: Multiple sequence alignment of yeast and human PCI subunits from proteasome lid, CSN and eIF3. Only the segments matched by the PROSITE PCI domain are shown. Conserved residues printed on black background were found in at last 50 % of ~60 PCI proteins of selected species, of which only yeast (sc) and human (hs) representatives are shown. Grey background was assigned to positions occupied by residues with similar physicochemical properties in at least 50 % of the sequences. Above the PCI alignment secondary structure prediction as calculated from JPred (Cuff et al., 1998) is presented. In these calculations sequences of eIF3k homologues were not included. Secondary structure elements of eIF3k as derived from PDB structure 1RZ4 are shown in a separate row. The abbreviations denote the following secondary structure types: E extended (sheet) and H helix. In addition, structural subdomain classification ('HAM', 'WH') as described in Wei et al. (Wei et al., 2004) and domain boundaries according to PCI profiles from PROSITE and Pfam are provided.


← HAM
|← PCI_DOMAIN
|← α-hairpin 1 →| |← α-hairpin 2
SEK_EIF3K
jpred_pci
EIF3K_HS 22
PSMD3_HS 278
PSMD6_HS 170
PSMD8_HS 44
PSMD11_HS 199
PSMD12_HS 214
PSMD13_HS 150
CSN1_HS 219
CSN2_HS 224
CSN3_HS 174
CSN4_HS 176
CSN8_HS 6
CSN12_HS 202
RPN3_SC 262
RPN5_SC 208
RPN6_SC 208
RPN7_SC 200
RPN9_SC 166
RPN12_SC 51
PCI8_SC 164
CSN12_SC 204
|← PCI PFAM HAM →|
→| |← α-hairpin 3 →| |← WH
SEK_EIF3K
jpred_pci
EIF3K_HS
PSMD3_HS
PSMD6_HS
PSMD8_HS
PSMD11_HS
PSMD12_HS
PSMD13_HS
CSN1_HS
CSN2_HS
CSN3_HS
CSN4_HS
CSN8_HS
CSN12_HS
RPN3_SC
RPN5_SC
RPN6_SC
RPN7_SC
RPN9_SC
RPN12_SC
PCI8_SC
CSN9_SC
CSN12_SC
PCI_DOMAIN →|
PCI PFAM →|
SEK_EIF3K
jpred_pci
EIF3K_HS
PSMD3_HS
PSMD6_HS
PSMD8_HS
PSMD11_HS
PSMD12_HS
PSMD13_HS
CSN1_HS
CSN2_HS
CSN3_HS
CSN4_HS
CSN8_HS
CSN12_HS
RPN3_SC
RPN5_SC
RPN6_SC
RPN7_SC
RPN9_SC
RPN12_SC
PCI8_SC
CSN9_SC
CSN12_SC

3.7.3.4 A structural model for the canonical PCI domain

The discovery of a PCI domain in eIF3k is of particular importance, as a three-dimensional structure of eIF3 has been solved recently (Wei et al., 2004). So far, no structural information on the PCI domain has been available, and a structural model for the canonical PCI domain based on the alignment shown in Figure 3-25 should allow interesting insights into the architecture of the PCI complexes.

A detailed analysis of the eIF3k structure (Wei et al., 2004) reveals a bipartite structure of two subdomains that are in close contact through a large inter-domain surface patch (Figure 3-26 (A)). The C-terminal half-domain is a globular α/β structure with an "αβααββ" arrangement. The three β-strands are very short and form an antiparallel sheet. The whole C-terminal part can be classified as a "winged helix" fold and thus is referred to as "WH-domain" (Wei et al., 2004). By contrast, the N-terminal half-domain is entirely helical with a core of six regularly-spaced helices that form three antiparallel helical hairpin elements. The resulting superhelix is reminiscent of the solenoids found in helical repeats such as HEAT, Armadillo and TPR. Somewhat unusual are the short 3_{10} helices that connect the consecutive α-hairpins. According to Wei et al. (Wei et al., 2004), the N-terminal half-domain resembles structurally mainly HEAT and Armadillo repeats, and thus the name "HAM-domain" was proposed. The bipartite structure of eIF3k is in good overall agreement with the secondary structure predictions for the single PCI domain families and also with the result of TPR-like helical repeats partially overlapping the PCI domain. It was therefore of special interest to make a detailed comparison of the eIF3k structure and the profile-guided alignment of the canonical PCI superfamily shown in Figure 3-25.

Within the N-terminal subdomain, the sequence conservation between the different PCI domain families is relatively poor and some aspects of the alignment shown in Figure 3-25 are not very reliable. Nevertheless, there is a good correspondence between the helices that build the α-hairpins of eIF3k and the uninterrupted sequence blocks in the PCI alignment. The gap-regions in the PCI alignment are typically caused by insertion events in selected PCI subfamilies. In no case, a deletion of one or more of the hairpin helices is observed. This finding suggests that the helical hairpin structure is conserved in most or all PCI domains. The constructed alignment and the derived secondary structure predictions suggest that the short 3_{10} helices that connect the helical hairpins in eIF3k are absent in most other PCI proteins. As mentioned in the previous paragraphs, there are several instances of subsignificant sequence similarity to TPR repeats found also *within* the N-terminal subdomain of the PCI domain. By contrast, no similarity to HEAT or Armadillo-repeats has been observed. Thus, the helical

hairpin structure of the N-terminal subdomain should be preferably interpreted as atypical TPR-like repeats rather than as the HEAT/Armadillo repeats suggested by Wei et al. (Wei et al., 2004).

The globular C-terminal subdomain (WH) is generally better conserved than the helical N-terminal domain and as a consequence, the part of the alignment covering this structural subdomain shown in Figure 3-25 is more reliable. The "αβααββ" arrangement of α- and β-regions is distributed over two large sequence blocks with a single major gap region between "αβα" and "αββ". As can be seen in Figure 3-25, no important secondary structure element is interrupted by a gap found in the PCI alignment. Like in the N-terminal subdomain, the WH portion shows a good concordance between the secondary structure predicted from the canonical PCI families and the structural elements of the eIF3k structure, apart from minor problems in predicting one of the very short β-strands.

Taken together, the comparison of the PCI alignment with the eIF3k structure shows that the two structures are clearly compatible and suggests that the canonical PCI domains will have an analogous bipartite fold similar to that shown in Figure 3-26. The prediction of TPR-like helical repeats N-terminal of the proper PCI domain suggests that they form an extension of the helical repeat region of the first PCI subdomain. The implications of this model for the overall PCI structure will be discussed in chapter 4.2.5.

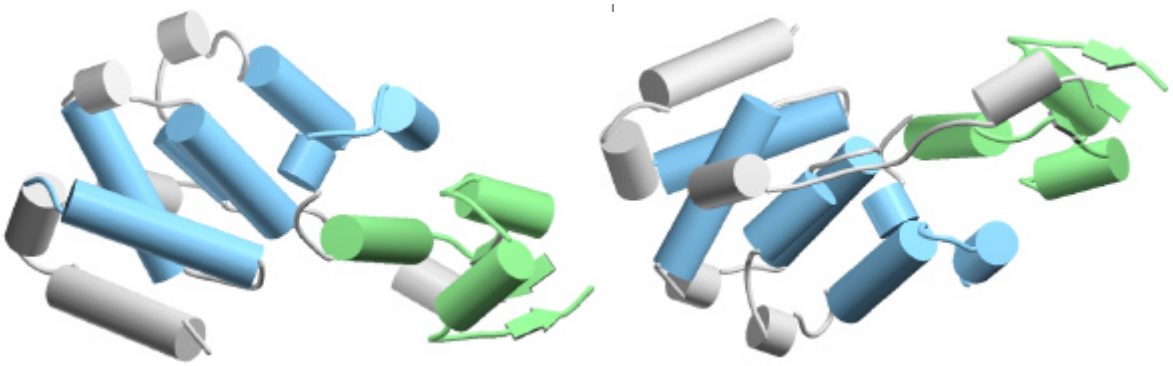


Figure 3-26 shows the overall structure of eIF3k from the PDB-entry 1RZ4 viewed from two sides (Wei et al., 2004) with β-strands and α-helices represented as ribbons and cylinders, respectively. Regions of the structure with sequence similarity to canonical PCI domain are rendered in colour. Regions belonging to the WH subdomain are shown in green, while conserved structure elements of the helical hairpin regions are shown in dark blue. The connection between β-strand 2 and 3 is not resolved and thus missing in 1RZ4. Other regions (extreme N- and C-termini, connecting helices between hairpins, unstructured regions) are shown in grey.

3.7.4 Other activators of the proteasome

3.7.4.1 PA28αβ, PA28γ and PA200

While the 19S regulatory particle ('RP') mediates Ub-dependent protein degradation by the proteasome, there exist at least three additional proteasomal activators that only stimulate the hydrolysis of peptides. These activators are generally referred to as PA28αβ, PA28γ and PA200 (Hendil, 1998, Tanahashi, 2000, Ustrell, 2002). Similar to the 19S RP, they can bind to each end of the proteasome, but mixed modes are possible as well, e.g. a 20S proteasome can associate with a 19S RP on one side, while the other one is bound to one of the other activators, PA28αβ or PA28γ. According to HUGO nomenclature the approved symbols for these genes are PSME1, PSME2 and PSME3, respectively (see Table 3-37). The corresponding genes are closely related and found in higher eukaryotes while homologous genes in yeast could not be detected. In contrast, human PA200 (or PSME4) has a yeast orthologue, Blm3/Blm10. While the binding of PA200 to the 20S proteasome has been shown, the physiological meaning remains elusive as discussed in more detail by Rechsteiner and Hill (Rechsteiner, 2005).

Table 3-37 Proteasomal activators. Human PA200 is orthologous to yeast BLM3/BLM10.

Proteasomal activators			
Yeast		Human	
Gene name	ORF	Gene name	Uniprot number
-	-	PSME1/PA28α/11S/REGα	Q06323
-	-	PSME2/PA28β/11S/REGβ	Q9UL46
-	-	PSME3/PA28γ/11S/REGγ	Q12920
BLM3/BLM10	YFL007W	PSME4/PA200	Q14997

3.7.5 Proteins involved in subunit synthesis and assembly of the 26S proteasome

3.7.5.1 Rpn4

The transcription of proteasomal subunit genes in yeast is positively regulated by the transcription factor Rpn4, which binds to an upstream activating sequence of these genes (Mannhaupt, 1999). As Rpn4 itself is degraded by the proteasome, the action of Rpn4 is regulated by means of a negative feedback loop, i.e. high proteasomal activity is connected to low Rpn4 levels resulting in low transcription of proteasomal subunits (Xie, 2001). Yeast Rpn4 carries a C2H2-type zinc finger at its C-terminus easily detectable across the fungal orthologues, while the sequence remainder is poorly conserved. No Rpn4 orthologues could be detected in

metazoan species in this work and the co-regulation of human proteasomal subunits remains elusive in general (see Table 3-38).

3.7.5.2 Ump1/hUMP1

Yeast Ump1 has been shown to assist 20S proteasome assembly and to play role in processing the initially inactive β-subunit precursors. For this function, Ump1 is localized in the central chamber of the 20S proteasome and becomes degraded after activation of the catalytic β-subunits (Ramos, 1998). A human orthologue was readily detectable in the database, hUMP1, which has already been classified as functional equivalent to Ump1 (see Table 3-38) (Burri, 2000, Witt, 2000).

Table 3-38 Proteins involved in proteasomal biogenesis.

Yeast Gene name	ORF	Human Gene name	Uniprot number
UMP1	YBR173C	hUMP1/POMP	Q9HB69
RPN4	YDL020C	-	-

3.7.5.3 Hsp90

The yeast homologues of the mammalian heat-shock protein Hsp90 have been reported to bind to the 26S proteasome and to play an important role for proteasome assembly as well as for its maintenance (see Table 3-39) (Imai, 2003).

3.7.6 Proteins involved in substrate delivery to the proteasome

3.7.6.1 Ub-like/UBA-adaptor proteins

At least the three yeast proteins Rad23, Dsk2 and Ddi1 have been shown to bind to the 26S proteasome and thereby mediate the recognition of ubiquitylated substrates (Elsasser, 2004, Elsasser et al., 2002, Saeki, 2002, Verma, 2004). These proteins are often considered as substrate delivery factors and share a similar domain topology consisting of an N-terminally located Ub-like domain followed by one or two UBA domains (see chapter 3.1.2.1). While the Ub-like domain binds to a proteasomal subunit, which is Rpn1 in the case of Rad23 (Elsasser et al., 2002), the UBA domain of the substrate delivery factors interacts with poly-ubiquitylated substrates. In human, there is more than one orthologue for each of these proteins, which all share a common domain topology with their yeast orthologues (see Table 3-4 and Table 3-39).

3.7.6.2 Ub-recognition components of the proteasome

In addition to substrate delivery factors, substrates to be degraded can be recognized by the proteasome directly. The yeast Rpn10 subunit of the base subcomplex acts as a receptor of ubiquitylated substrates, probably mediated by the UIM of Rpn10 (Elsasser et al., 2004). This subunit is found as a single copy in both yeast and human (Table 3-34). Another subunit of the proteasome involved in substrate binding is human PSMC3/S6a, an AAA ATPase subunit of the base (Lam et al., 2002). This protein has a clear yeast orthologue, termed Rpt5.

3.7.6.3 Hsp90/Hsp70/BAG1/CHIP

The human chaperone Hsp90 or HSPCA has already been assigned a role in the UPS as substrate specificity factor for the E3 ligase CHIP (Cyr et al., 2002). It is also discussed as acting in parallel to PA28αβ in the transfer of peptides from the proteasome to the class-I loading complex located in the membrane of the ER (Yamano, 2002). In yeast, two nearly identical homologues of Hsp90 exist (96% similarity), which are redundant in function and differ only in their expression pattern. The two yeast proteins are referred to as Hsc82 (Ymr186w) and Hsp82 (Ypl240c). Both genes were found to be orthologous to human HSPCA and its close relative, HSPCB/HSP90B, which shares 86% similarity with HSPCA (see Table 3-39). Besides these two human proteins, the human genome encodes two additional members of the Hsp90 family, TRA1 (P14625) and TRAP1/HSP75 (Q12931). Whether the latter two proteins play a role for proteasomal proteolysis is still an open question.

BAG1 is a co-chaperone that binds to Hsp70 via its BAG domain (Bcl2-associated athanogene) resulting in subsequent release of Hsp70 bound substrates destined for degradation (Bimston, 1998). The Ub-ligase CHIP, which is associated with the chaperones Hsp70 and Hsp90, poly-ubiquitylates BAG1 in an unusual Lys-11 linkage (Alberti, 2002). This type of linkage does not induce BAG1 degradation, but stimulates its association with the proteasome. However, the BAG1 Ub-like domain also seems to be involved in proteasome binding (Luders, 2000). Taken together, BAG1 acts in the delivery of Hsp70-bound substrates to the proteasome. Despite extensive profile-based searches, no BAG1 or CHIP homologue in yeast could be detected (see Table 3-39).

3.7.6.4 Cdc48/VCP/p97

The AAA ATPase Cdc48/VCP/p97 is another factor for substrate delivery to the proteasome. It forms hexameric ring-shaped complexes, which recruit several cofactors like p47 or Ufd1 needed for substrate proteolysis at the ER by the ERAD pathway (Richly, 2005). For all proteins so far known that have a role in Cdc48 mediated proteasomal delivery, clear orthologues exist between yeast and human (see Table 3-39).

Table 3-39 Proteins involved in substrate delivery to the proteasome.

Yeast		
Gene name	**ORF**	**Comments**
OAZ1	YPL052W	partial ORF; Ub-independent
UBR1	YGR184C	N-end-rule
CDC48	YDL126C	CDC48/p97/VCP
UFD1	YGR048W	CDC48/p97/VCP
NPL4	YBR170C	CDC48/p97/VCP
SHP1	YBL058W	CDC48/p97/VCP
RAD23	YEL037C	UBA/Ub-like adaptor
DSK2	YMR276W	UBA/Ub-like adaptor
DDI1	YER143W	UBA/Ub-like adaptor
Human		
Gene name	**Uniprot number**	**Comments**
OAZ1	P54368	Ub-independent
UBR1	Q8IWV7	N-end-rule
UBR2	Q8IWV8	N-end-rule
UFD1	Q92890	CDC48/p97/VCP
VCP/p97	P55072	CDC48/p97/VCP
NPL4	Q8TAT6	CDC48/p97/VCP
NSFL1C/p47	Q9UNZ2	CDC48/p97/VCP
RAD23A	P54725	UBA/Ub-like adaptor
RAD23B	P54727	UBA/Ub-like adaptor
UBQLN1/PLIC1	Q9UMX0	UBA/Ub-like adaptor
UBQLN2/PLIC2	Q9UHD9	UBA/Ub-like adaptor
UBQLN3	Q9H347	UBA/Ub-like adaptor
UBQLN4/UBIN	Q9NRR5	UBA/Ub-like adaptor
DDI1	Q8WTS3	UBA/Ub-like adaptor
DDI2	Q7RTZ0	UBA/Ub-like adaptor
HSPA1A/Hsp70	Q5SP17	Hsp90/Hsp70/CHIP/BAG1
HSPCA/HSP90A	P07900	Hsp90/Hsp70/CHIP/BAG1
HSPCB/HSP90B	P08238	Hsp90/Hsp70/CHIP/BAG1
BAG1	Q99933	Hsp90/Hsp70/CHIP/BAG1
CHIP	Q9UNE7	Hsp90/Hsp70/CHIP/BAG1
Assignment of orthologues		
Yeast	**Human**	**Comments**
OAZ1	OAZ1	Ub-independent
UBR1, UBR2	UBR1, UBR2, UBR1L1	N-end-rule (not sure for UBR1L1)
UFD1	UFD1	CDC48/p97/VCP
CDC48	VCP/p97	CDC48/p97/VCP
NPL4	NPL4	CDC48/p97/VCP
SHP1	NSFL1C/p47	CDC48/p97/VCP
RAD23	RAD23A, RAD23B	UBA/Ub-like adaptor
DSK2	UBQLN1-4	UBA/Ub-like adaptor
DDI1	DDI1, DDI2	UBA/Ub-like adaptor
HSP82, HSC82	HSPCA, HSPCB	Hsp90/Hsp70/CHIP/BAG1

3.7.6.5 Antizyme: a model for Ub-independent proteasomal targeting

Note: the following work has been published as part of *"Polyamines regulate their synthesis by inducing expression and blocking degradation of ODC antizyme."* by Palanimurugan et. al. in EMBO J. 2004 Dec 8;23(24):4857-67 (Palanimurugan, 2004).

Antizyme mediates Ub-independent degradation of ODC by the proteasome

Antizyme ('anti-enzyme for ornithine decarboxylase') was first discovered as inhibitor of ODC (ornithine decarboxylase), a key regulator of polyamine biosynthesis. Antizyme mediates the degradation of ODC by the proteasome in a Ub-independent manner (Murakami, 1992). Noticeably, antizyme is not degraded during this process and recycled instead allowing subsequent rounds of antizyme-mediated ODC degradation. The proper regulation of ODC and therefore the regulation of the polyamine levels is much more complicated, as polyamines themselves increase the antizyme level in a negative feedback loop. The detailed mechanism relies on a frameshifting event induced by polyamines during the translation of antizyme mRNA. Without the frameshift, an in-frame stop codon (mostly TGA) causes the expression of a truncated and inactive variant of antizyme (see Figure 3-29) (Matsufuji, 1995).

A yeast antizyme has not been identified so far

Antizyme is a widespread protein known in a lot of eukaryotic organisms ranging from fungi, insects, nematodes to higher organised species like the pufferfish or mammals. The more complex the organism, the more antizyme homologues can be found in its genome, e.g. there are at least four human antizyme homologues. Although detectable in fungi like *S. pombe*, the corresponding antizyme in *S. cerevisiae* and the closely related Hemiascomycetes is undiscovered so far (Zhu, 2000). But experimental data in the field of ODC regulation in yeast exist leading to the postulation of a yeast antizyme (Gandre, 2002, Toth, 1999). To test for an existing yeast antizyme, several sequence-based analyses were performed, but only profile-based methods rather than pairwise methods succeeded in reporting significant candidates. Neither BLAST searches starting from ascomycetes like *P. carinii* or *S. pombe* antizyme did return significant matches in public Saccharomyces databases above threshold, nor could the HMM derived from the fungal antizyme family (from Pfam version 12.0) detect any potential antizyme homologues in yeast. Therefore, new antizyme-based profiles were constructed in order to create a more sensitive means for finding a yeast antizyme.

Defining an antizyme region as starting point for profile searches

A careful consideration of the antizyme family alignment reveals two well conserved regions separated from each other by a divergent linker sequence of variable length depending

on the species. This domain structure seems to be maintained through all obvious members of this family, which comprises more than 40 proteins until now. The conserved N-terminal segment (D1) is centred around the frameshifting site in all antizymes and is approximately 50 residues away from the start methionine. The highly conserved C-terminus (D2), which presumably carries the structural requirements for binding ODC (Li, 1994), was chosen as a starting point for the search for yeast antizyme.

As the gene tree of the known antizymes largely reflects the 18S rRNA derived evolutionary history, the yeast antizyme was expected to adopt a similar behaviour and to cluster together with *S. pombe* or *P. carinii* (Berbee, 2000) in a tree derived from antizyme sequences. For this reason, an MSA of antizyme sequences from fungi of the Schizosaccharomyces clade was calculated and trimmed to the D2 region described above. In detail, the starting profile was built from three protein sequences derived from *S. pombe* (Q9USQ5), *S. japonicus* (Q9HFU9) and *S. octosporus* (Q9HFU8).

S. cerevisiae ORF Ypl052w is a putative antizyme homologue

A subsequent search in the sequence database yielded significant matches in obvious homologues from *P. carinii* (a yeast from the Schizosaccharomyces clade), *B. cinerea* (another Ascomycete) and insects like *A. gambiae*. These sequences were integrated into the existing alignment covering the D2 domain. This alignment in turn served as a starting point for a new round of profile construction and database search. The profile from round two retrieved additional antizymes from fungi like *F. gramineum, E. nidulans* and *N. crassa* with significant scores and an extra vertebrate sequence from pufferfish. No members of obvious non-antizyme families were returned at this point of the search.

After several rounds of profile constructions and searches the complete set of known antizymes from vertebrates was collected in addition to antizymes from nematodes and some fungal sequences from *C. neoformans, U. maydis* and *S. kluyveromyces*, the latter a member of the Saccharomycetaceae clade and therefore a close relative of *S. cerevisiae*. During the next profile iteration an uncharacterised *S. cerevisiae* ORF named Ypl052w and its obvious orthologues from *S. castelli, S. kudriavzevii, S. mikatae, S. bayanus* and from the two recently published genomes of *A. gossypii* and *K. waltii* (Dietrich, 2004, Kellis et al., 2004) gained significant scores, rendering them as highly potential candidates for the missing antizyme orthologues in these fungi.

Reverse profile searches were performed relying on Ypl052w and its orthologous sequences from the other Saccharomyces fungi as initial input in order to validate them as proper antizyme homologues. The re-detection of known antizymes turned out to be

challenging, as the diversity within the Saccharomyces subfamily was to low to confer enough sensitivity on the profile at this point. Even with clear *A. gossypii* and *K. waltii* orthologues of the potential yeast antizyme available, the profile of the reverse iteration converged at an early stage comprising only the close homologues of yeast antizyme. Nevertheless, known antizymes from *S. pombe* and *P. carinii* appeared as potential matches scarcely below the significance threshold and asymmetric results from profiles with distinct starting points are an everyday observation.

Characterizing the yeast antizyme sequence

The genomic sequence around ORF Ypl052w exhibits some features similar to that found in known antizyme mRNAs, e.g. an alternative start codon exists ~300 bases upstream of the YPL052W start codon, followed by an in-frame stop codon at bases 208 to 210 (see Figure 3-27). This putative open reading frame was termed ORF1. Upon a predicted +1 ribosomal frameshifting event skipping the 't' of the ORF1 stop codon, translation continues to base 877 (ORF2). While the end of ORF2 and YPL052W are identical, ORF2 is longer at its 3'-end due to a so far assumed start codon for YPL052W at base 274. Translating ORF1 and ORF2 while incorporating a +1 frameshifting event at the in-frame stop codon leads to a putative protein sequence, whose N-terminal region (ORF1) exhibits homology to the D1 motif in antizymes. Especially the residues in immediate vicinity to the putative frameshifting site could be aligned to well conserved positions in the antizyme family. This observation is valid for all orthologues from the Saccharomyces branch indicating that the frameshifting site is likely present in these sequences as well (see Figure 3-30). Concerning the two NES regions, it was not possible to affirm the first one, which should be located upstream of D1, but the second one could be detected as part of the D2 domain.

In human antizyme, the sequence 3'-wards of the shifty stop codon has been described as a pseudoknot constituting stretch (Namy, 2004). However, the corresponding region in yeast is devoid of any sequence similarity to the human mRNA sequence and an RNA secondary structure prediction detected only a weak folding propensity ~20 bases 3'-wards of the frameshifting site (data not shown). In contrast, *S. pombe* antizyme mRNA probably has a defined secondary structure behind the shifty stop codon. Considering the protein sequence of this trans-frameshifting region, the assignment of sequence similarity using pairwise alignments between the known antizyme family and yeast was extremely difficult.

The D2 domain appears as a well conserved segment between the potential Saccharomyces antizymes. Corroborating the observations from the reverse profile search, the conservation between the Saccharomyces proteins and the known antizymes appears somewhat

worse (Figure 3-30). In general, the D2 domain consists of five primary sequence blocks separated by short divergent linkers in all antizymes. Antizymes of *S. cerevisiae* and its relatives are somewhat exceptional as they have large stretches of 20 to 30 residues inserted between the second and the third block. This insertion may be interpreted as a linker that connects a region lacking a predictable secondary structure and a likely αβα-fold. In the latter, each secondary structure element corresponds to one of the three remaining sequence blocks in the alignment. A comparable insertion is not observed in antizymes of higher eukaryotes and might be the reason for the difficulties encountered in detecting the Saccharomyces antizymes and vice versa, finding the known antizymes with profiles starting from the Saccharomyces sequences. It should be mentioned that no antizyme orthologue in *Candida albicans* could be detected, may be due to insufficient sequence data.

To summarize, a budding yeast antizyme is predicted here, which has not been found in a previous search for the yeast antizyme (Zhu et al., 2000). The coding region is located on chromosome XVI and includes ORF1 and ORF2, which contains YPL052w, as described above. If the prediction of yeast antizyme is correct, synthesis of this protein would also require a ribosomal frameshifting event as it is the case for all established antizymes. The obvious sequence similarity between the putative Saccharomyces antizymes and the established antizymes from other eukaryotes, the conservation of the domain structure and the presence of an in-frame stop-codon point at a common evolutionary origin and function between these proteins. Particularly, the substructure of D2 and the identical secondary structure prediction of an αβα-fold at the tail of D2 were in strikingly good agreement with known antizyme features and made the biochemical validation of the potential yeast antizyme promising. Indeed, the predicted yeast antizyme was experimentally validated by Palanimurugan et al., who have clearly demonstrated its influence on ODC degradation, the need of a frameshifting event for proper expression and the influence of polyamines on the frameshifting event (Palanimurugan et al., 2004).

Figure 3-27 Genomic sequence of the budding yeast OAZ1 locus together with the primary sequence of the corresponding polypeptide. Below each amino acid the corresponding codon is shown except for the frameshifting site (bold letters), where the shifted 't' is depicted above the 'gac' codon translated in full-length Oaz1. ORF1 encompasses the OAZ1 gene from base 1 to base 210 ending with an 'tga' (bold). Skipping the 't' of this codon generates the 'gac' codon, which constitutes the beginning of ORF2 (base 209 to 880). An asterisk marks the Ypl052w (ORF2) start codon.

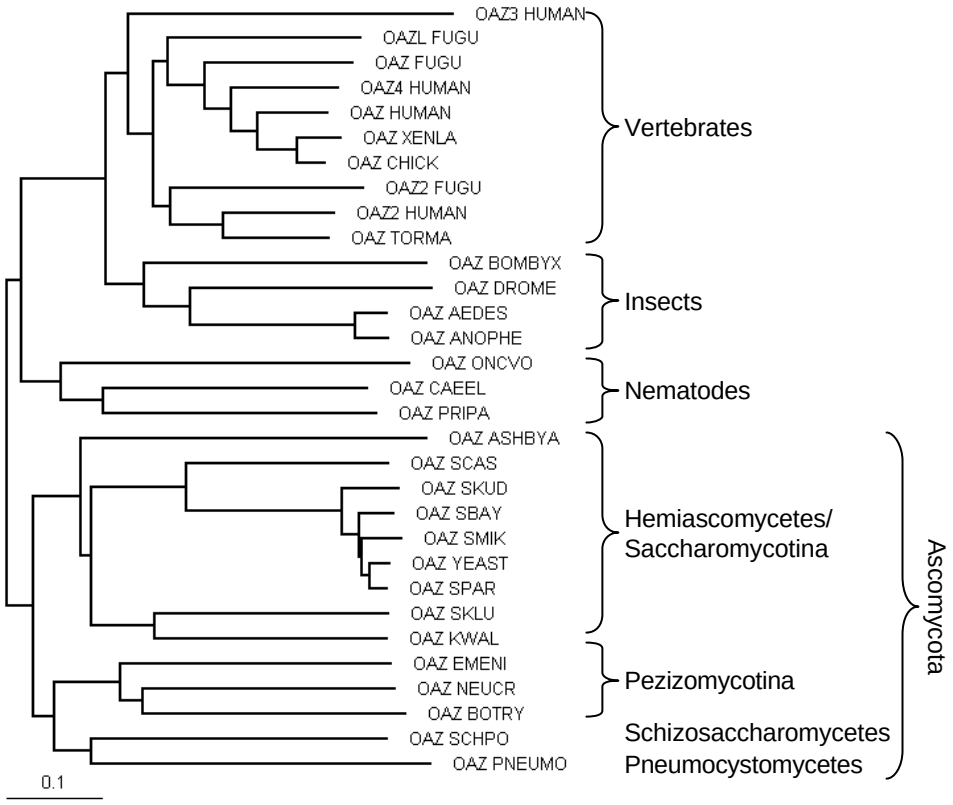


Figure 3-28 Neighbour-joining dendrogram of selected antizyme sequences. Only gap-less and reliable alignment positions were considered for the tree construction. Species names are abbreviated as in the multiple alignment of antizyme proteins sequences except for TORMA, *Torpedo marmorata*; BOMBYX, *Bombyx mori*; FUGU, *Takifugu rubripes*; CHICK *Gallus gallus*; AEDES, *Aedes aegypti*; ANOPHE, *Anopheles gambiae*; ONCVO, *Onchocerca volvulus*; PRIPA, *Pristionchus pacificus*; ASHBYA, *Ashbya gossypii*; SKUD, *Saccharomyces kudriavzevii*. Major taxonomic ranks are indicated at the right border of the figure.

Figure 3-29 This figure contains a DNA alignment covering the frameshifting region of established antizymes (lower panel) and previously unrecognised orthologues from Hemiascomycetes in the upper panel. The marked base is read through during +1 slippage. The last column shows EMBL accession numbers for the corresponding DNA sequence if available. Hemiascomycete DNA sequences were found as homologues of YPL052w in the Saccharomyces genome database (SGD). Species abbreviations are given in Figure 3-28.

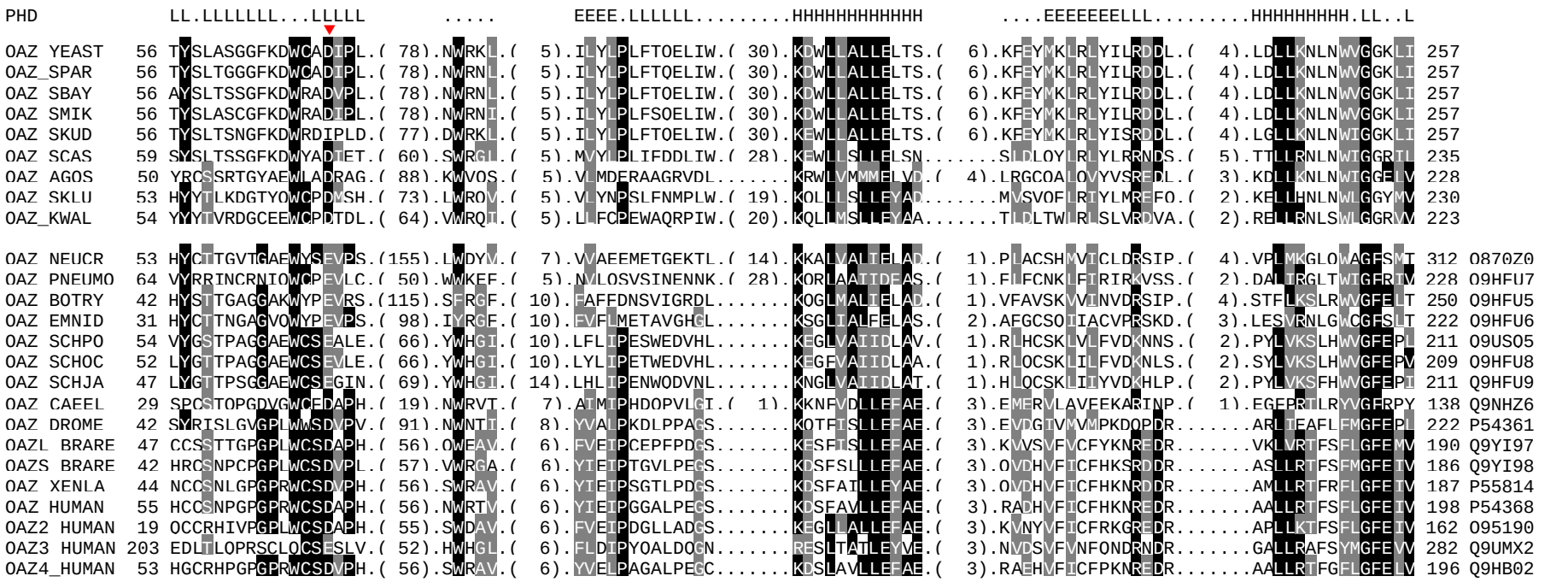


Figure 3-30 This figure shows the protein alignment of representative antizymes covering metazoan and fungal organisms. The alignment is subdivided into two segments with the established antizymes in the lower panel and newly discovered orthologues from hemiascomycetes in the upper one. Alignment positions conserved are printed on black background or accordingly positions assigned to amino acids with similar physicochemical properties are shaded in grey if supported by 50% or more of all antizyme family members. A red triangle indicates the position of the frameshifting site. In the top row the secondary structure as calculated by PHD is presented; the abbreviations denote following secondary structure types: E extended (β-sheet), A α-helix and L loop; only positions with an expected average accuracy > 82% were considered. Sequence coordinates are provided on both ends of the alignment. Last column depicts the Uniprot accession number if available. Species names are abbreviated: SPAR, *Saccharomyces paracelsus*; SBAY, *Saccharomyces bayanus*; SMIK, *Saccharomyces mikatae*; SCAS, *Saccharomyces castellii*; AGOS, *Ashbya gossypii*; SKLU, *Saccharomyces kluyveri*; KWAL, *Kluyveromyces waltii*; NEUCR, *Neurospora crassa*; PNEUMO, *Pneumocystis carinii*; BOTRY, *Botrytis cinerea*; EMNID, *Emericella nidulans*; SCHPO, *Schizosaccharomyces pombe*; SCHOC *Schizosaccharomyces octosporus*; SCHJA, *Schizosaccharomyces japonicus*; CAEEL, *Caenorhabditis elegans*; DROME, *Drosophila melanogaster*; BRARE, *Brachydanio rerio*; XENLA, *Xenopus laevis*.

3.7.6.6 Ubr1/ClpS: a common domain in the N-end-rule-pathway of eukaryotes and bacteria

So far, well established homology domains or protein families with a role in substrate delivery to the proteasome have been dealt with. In the following, a novel homology domain with a putative role in substrate delivery to chambered multi-subunit proteases will be described.

Ubr1 interacts with the proteasome

The RING-finger-type Ub-ligase Ubr1 (N-recognin) is part of the N-end-rule pathway, together with Ubc2 as E2 and Uba1 as E1. The N-end-rule-pathway is used for the proteasome dependent degradation of proteins with a destabilizing N-terminal residue. Within this pathway, Ubr1 has multiple functions, which are recognition of the target, ubiquitylation and probably the delivery of the substrate to the proteasome. The latter is suggested by the finding that yeast Ubr1 binds to the Rpn2, Rpt1 and Rpt6 subunits of the 19S regulatory particle of the 26S proteasome (Xie, 2000). The sites of Ubr1 interacting with the proteasome have not been mapped so far. Therefore, a detailed sequence analysis of Ubr1 has been carried out in order to predict a possible proteasome-binding site.

Finding a putative interaction domain

Ubr1 is a large multi-domain protein of ~225 kDa in yeast, but only small portions of Ubr1's primary sequence are covered by known domains. At the N-terminus of yeast Ubr1, a putative zinc finger domain ('ZF_UBR1') was found that is needed for recognition of certain types of substrates (Kwon, 1998). A RING finger domain is located ~1000 residues downstream and has an essential role in the formation of substrate-linked Ub-chains (Xie, 1999). Besides the two zinc fingers, an eight residue long BRR-motif (for 'basic residue rich') is located directly upstream of the RING finger and has been reported to be important for Ubc2-binding (Xie et al., 1999). Obviously, much of the remaining primary sequence is left without functional annotation. At the same time, there is sufficient conservation between fungal and mammalian Ubr1 orthologues to allow profile construction for untouched regions in order to reveal new homology domains.

Common homology domain in Ubr1 and ClpS

In one of the attempts to derive novel homology domains of Ubr1, a sequence segment adjacent to the C-terminus of the ZF_UBR1-domain was used for profile construction. The latter started from a two-sequence-alignment of the corresponding sequences from yeast and *C. albicans* Ubr1. After two iteration cycles, all eukaryotic Ubr1 homologues available in the

databases were already detected with significant scores. In addition, CT2237 (Uniprot Q8KAC6) from the cyanobacterium *Chlorobium tepidum* was matched in a highly significant manner (N=14.5, p=10e-5.6) and almost the whole sequence was involved in this match (~100 residues). To validate this finding, a control profile was constructed from CT2237 and its homologues. A subsequent profile search revealed a large set of bacterial proteins of similar size. The detected bacterial family is generally referred to as ClpS family. In following iteration cycles, members of the original Ubr1 family were detected, thus verifying the homology relationship between the considered Ubr1 region and the bacterial protein family. The homologous region spans nearly the whole sequence in most bacterial proteins. Therefore, the precise domain boundaries could be derived easily, which is not always an easy task (see chapter 3.7.2). The N-terminal half was found to be highly conserved and to contain mainly hydrophobic residues (see Figure 3-31). The C-terminal half is more divergent, but contains several charged residues found in many sequences. It should be mentioned that during this work, similar findings were reported by Lupas and Koretke (Lupas, 2003).

ClpS is a substrate modulator of ClpAP with a known structure

Most bacteria encode one ClpS homologue, which is typically found in an operon together with ClpA (Dougan, 2002). ClpA is an AAA-ATPase type chaperone forming complexes together with the protease ClpP. The resulting multimeric ClpAP complex is a chambered protease consisting of two heptameric ClpP rings with one or two ClpA hexamers attached. This complex degrades proteins in an ATP-dependent manner in the bacterial cytosol similar to the eukaryotic proteasome. Dougan et al. have shown that the substrate specificity of ClpAP upon binding to ClpS changed from SsrA-tagged proteins and ClpA itself to protein aggregates. Thereby, ClpS has been assigned the role of a substrate modulator for the ClpAP proteolytic complex. ClpS directly interacts with the N-terminus of ClpA and stoichiometric amounts of ClpS molecules are needed to make the substrate specificity switch of an individual ClpAP complex complete.

ClpS has a conical shape and consists of three α-helices facing an antiparallel β-sheet (Zeth, 2002). In Figure 3-32, *E. coli* ClpS is shown in complex with the N-terminal domain of ClpA. The residues conserved between the ClpS domain of Ubr1 and ClpS from *E. Coli* are mostly located in the hydrophobic core and therefore likely play a structural role, i.e. they stabilize the fold of the ClpS domain. Simultaneously, there is an excellent correspondence between the secondary structure of the *E. coli* ClpS and that predicted for the ClpS domain in Ubr1 homologues (see Figure 3-31). These findings suggest that the 3D structure of the ClpS domain is conserved between bacteria and eukaryotes.

ClpS and Ubr1 share residues essential for substrate binding

ClpS may act as an adaptor recruiting specific substrates to the ClpAP (Dougan et al., 2002). Interestingly, residues located in the ClpS domain of yeast Ubr1 seem to be essential for this Ubr1 homologue to bind to type-2 substrates of the N-end-rule pathway (see Figure 3-31) (Kwon et al., 1998). Moreover, the homologous residues of *E. coli* ClpS are located opposite to the ClpA binding site leaving space for additional interactions partners, e.g. substrates. From these observations it may be hypothesized that the function of the ClpS domain to bind substrates may be conserved between bacteria and eukaryotes, while its binding to a chambered protease is different or not present at all in eukaryotes. Probably, alternative regions of the large Ubr1 carry out the proteasome binding, while the ClpS domain only exerts a rudimentary function as substrate binding domain. At least in this regard, the ClpS domain would link both the bacterial and the eukaryotic N-end-rule pathway.

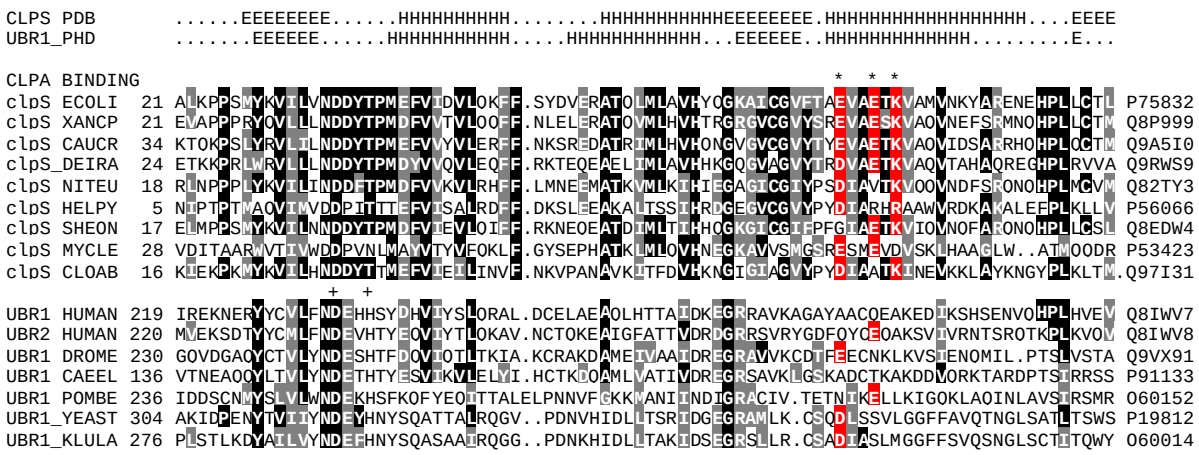
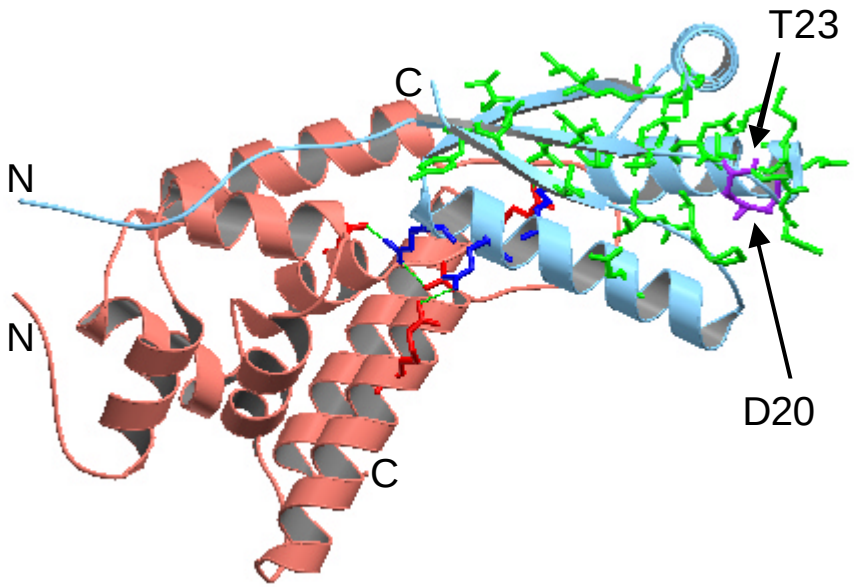


Figure 3-31 The ClpS domain. The alignment shows the ClpS domain of bacterial ClpS proteins and eukaryotic Ubr1 homologues. In the top rows, the secondary structure as derived from the ClpS structure of *E. coli* (pdb:1LZW) or as predicted for Ubr1 by JPred is shown, respectively. Conserved residues are printed on black background, and positions assigned to amino acids with similar physicochemical properties are shaded in grey if supported by at least 50% of all available ClpS domain containing proteins (not all are shown). Amino acid residues that are part of the ClpA-binding interface are indicated with asterisks and highlighted in red instead of black if conserved. A '+' is found above the position that has been reported as essential for type-2 substrate binding in yeast Ubr1 (Kwon et al., 1998). Species abbreviations are as follows: ECOLI, *E. coli*; XANCP, *X. campestris*; CAUCR, *C. crescentus*; DEIRA, *D. radiodurans*; NITEU, *N. europaea*; HELPY, *H. pylori*; SHEON, *S. oneidensis*; MYCLE, *M. leprae*; CLOAB, *C. acetobutylicum*; DROME, *D. melanogaster*; CAEEL, *C. elegans*; POMBE, *S. pombe*; YEAST, *S. cerevisiae*; KLULA, *K. lactis*.

A



B

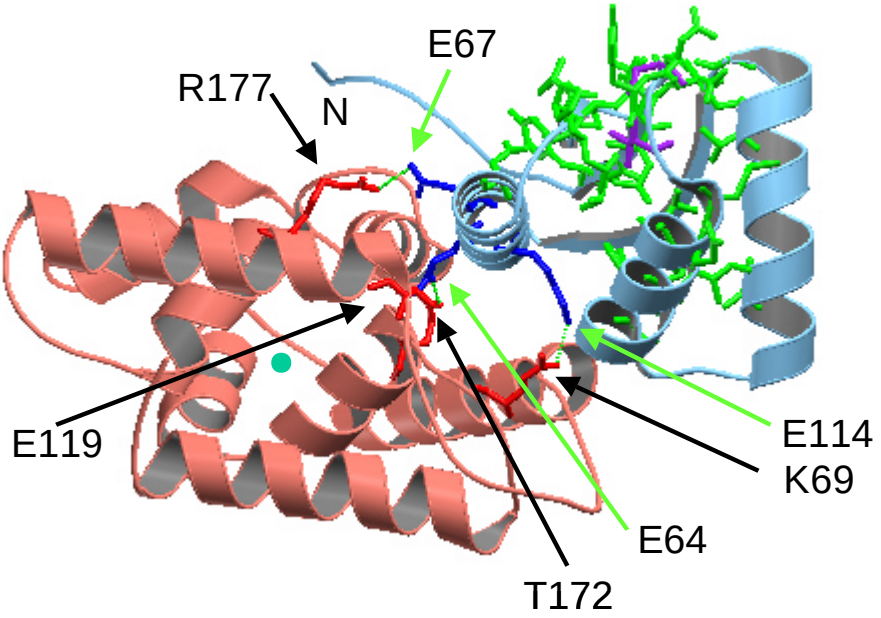


Figure 3-32 Structure of ClpS in complex with the N-terminal domain of ClpA. (A), a ribbon representation of ClpS in blue and ClpA N-terminal domain in red. ClpS adopts a conically shaped fold with an elongated N-terminal tail. Conserved residues between bacterial ClpS sequences and human Ubr1 are coloured in green. Most conserved residues point to the interior or stabilize secondary structure element such as the β-sheet. Purple residues are conserved and simultaneously essential for substrate binding in yeast Ubr1. Blue and red residues mediate the ClpS-ClpA interaction. Hydrogen bonds are depicted in dotted green lines. (B), same as A, but focus is now on the interaction interface and the view is along the long ClpS α-helix harbouring the three residues essential for binding to ClpA. Obviously, none of the conserved residues is in range of ClpA residues. A green dot represents the two-fold symmetry axis of the N-terminal ClpA domain.

3.7.7 Physiological proteasome inhibitors

3.7.7.1 PI31

PI31 is a negative regulator of immunoproteasome precursor maturation and, if overexpressed, abrogates the presentation of MHC class I antigens (Zaiss, 2002). Moreover, PI31 blocks the activation of 20S proteasomes by PA28 (Zaiss, 1999). So far, PI31 has only been reported in higher eukaryotes, but not in yeast. Therefore, profile-based searches were carried out covering distinct segments of PI31. Indeed, profiles constructed from a C-terminal stretch of PI31 (PI31_CTERM) detected the yeast ORF Ycr076c, though with weak significance. Searching the databases with a control-profile generated from Ycr076c, PI31 homologues were retrieved with plant homologues yielding better scores. However, the similarity between fungal orthologues of Ycr076c and animal PI31 homologues remains weak and experimental data are needed to validate this finding. Interestingly, the human F-box protein FBXO7 (Q9Y3I1) appeared to be related to PI31, as both share the PI31_CTERM as well as second domain, called PI31_NTERM.

4 Discussion

4.1 Cataloguing the UPS with the profile technique

Ubiquitin and related modifiers participate in a variety of different cellular processes

Ubiquitin is the founding member of a family of proteinogenic post-translational modifiers, which share a very similar three-dimensional structure. The original function of Ub has been described as marking substrates for proteasomal degradation. However, this function is just one of many roles Ub plays in the cell, as ubiquitylation of substrates has also been reported to regulate DNA damage repair, endocytosis, gene expression, chromatin structure and also lysosomal/vacuolar protein degradation (Schwartz, 2003). This variety of processes becomes even larger when additional Ub-like modifiers are taken into account, as they act as analogous signals to Ub when conjugated to substrates.

Variations on the Ub signal

One of the crucial aspects of Ub biology is that ubiquitylation can convey different signals depending on the place of substrate attachment, the chain length and the linkage type (Hicke, 2001, Weissman, 2001). Possibly, the number of Ub molecules attached to distinct sites in the substrate plays a role as well. For the selection of the appropriate type of the Ub signal, the ubiquitylation and deubiquitylation machinery must be strictly regulated.

The UPS is highly complex

The complexity of the UPS can already be seen by the high number of proteins involved. Another characteristic for the complexity is the occurrence of concerted mechanisms, e.g. in the ubiquitylation cascade or the proteasomal degradation step. Moreover, the specific substrate ubiquitylation as well as editing the Ub-signal of ubiquitylated substrates contribute additional regulatory levels to the UPS and therefore enhance its complexity.

4.1.1 Functional domains help structuring the UPS

In the literature, proteins relevant to the UPS have already been divided into certain functional classes. For example, there are classes covering proteins of the ubiquitylation machinery, proteasomal components, adaptors or deubiquitylating enzymes. By bioinformatical means, these protein classes have been further subclassified in the past. Usually, the concept of homology domains in combination with appropriate computational tools has been widely applied, because homology domains are often directly linked to a certain function of the

proteins they are found in. Besides the classification of UPS proteins, homology domain-based approaches have served as useful means in identifying new components of the UPS or related systems in the past. One reason for homology domains being successful in structuring the UPS and finding new components is the fact that the UPS associated homology domains are widely used throughout all eukaryotic realms of life and in certain cases can be traced back even to bacteria (Furukawa, 2000, Lupas et al., 2003). Another reason may be the advent of the sensitive profile technique that made major contributions to this field (Bai et al., 1996, Hofmann et al., 1996, Hofmann et al., 1998, Hofmann et al., 2001).

4.1.2 The UPS and related systems resemble intracellular signal transduction pathways

Intracellular signal transduction pathways comprise three basic steps: signal generation, signal recognition by specialized recognition domains and signal removal. Classic examples are kinase phosphorylation cascades (Bhagwat, 1999) and the apoptosis cascade involving specific adaptor proteins and caspases (Reed, 2004). For comparison, the UPS shows analogous steps such as ubiquitylation, ubiquitin recognition and deubiquitylation. The UPS therefore resembles signalling pathways. The same holds true for Ub-related modifiers such as SUMO and their corresponding conjugation and recognition mechanisms (Dohmen, 2004). The UPS and Ub-like pathways also resemble each other in the homology domains involved. As a consequence, the presence of a particular type of homology domain often does not allow to decide whether a protein belongs to the UPS or a related system. For example, at least four members of the human ULP family are established SUMO proteases, suggesting a general desumoylating activity of the entire ULP family. However, work of Gan-Erdene et al. shows that SENP8/Den1 is active against NEDD8 rather than SUMO (Gan-Erdene et al., 2003). On the other hand, pathways may overlap, as seen for UBE2E2/UCH8, which conjugates Ub as well as ISG15 (Zhao et al., 2004).

4.1.3 Benefits from cataloguing the UPS

In this work, the most important UPS associated homology domains known from literature have been re-analyzed in detail and used to set up a catalogue of proteins relevant to the UPS and its related systems from human and yeast. Amongst other things, these data enable studies on the evolutionary origin of the UPS. As a special focus of this work has been on the

assignment of orthologues, information may be transferred from the well-studied yeast UPS to the human system or vice versa.

4.1.4 Limitations of existing sequence profile collections

For searching the sequence databases for established and novel members of the UPS, the available profile-HMMs from Pfam and sequence profiles from PROSITE were employed in a first step. While comparing sets of retrieved proteins from profile searches with a Pfam profile-HMM and its corresponding PROSITE profile, often proteins were only found by one of the two methods. The reason for the different sensitivities of the HMMs and profiles used may be due to the incorporation of distinct subfamilies, or because of a different update status. In addition to the different sensitivities, Pfam and PROSITE often also use different domain boundaries for identical homology domains. An example for this behaviour has been described in the PCI domain in chapter 3.7.3. Several profiles have difficulties in discriminating between distinct protein families, e.g. the PHD finger and the RING finger profiles, which detected overlapping protein sets (see chapter 3.4.1.1). For that reason, the available profiles were not always sufficient to catalogue the UPS components. As a consequence, subfamily-specific profiles had to be generated for a lot of protein families, e.g. for the RING finger family. In other cases, the established profiles were absolutely appropriate, like the profiles derived from E1-, E2-, HECT-, UIM-proteins and some DUB families.

For constructing profiles within this work, the generalized profile technique was used because for most protein families the generalized profile technique exhibits a better sensitivity than profile HMMs, which often suffer from overfitting (Palanimurugan et al., 2004, Zhu et al., 2000). One problem of the profile technique is the danger of missing outliers or small families, even if they were originally being part of the seed MSA. These protein sequences run the risk of not being re-detected during the iterative profile improvement steps.

4.2 Revised and novel homology domains

In the following paragraphs, I will discuss some interesting examples where a revised profile was able to detect further UPS-relevant homology domains in novel proteins. Also, sequence analysis conducted on established UPS components provided information on new homology domains and motifs, for which a general role in the UPS is possible.

4.2.1 Novel Ub-like domains in E1s and their functional role

The recently reported 3D structures of the human E1 enzymes UBE1C/UBA3 and UBLE1B/UBA2 clearly show the presence of Ub-fold domains in those two proteins. However, neither Pfam nor PROSITE profiles were able to detect these domains. Profiles generated in this work could detect similar domains in additional E1s. However, this profile-based analysis also suggested that the sequence similarity between the E1 Ub-fold domain and the classic Ub-like domain is extremely weak.

The role of the Ub-fold in E1 proteins might be that of a binding site, e.g. for E2s. In a recent report, Huang et al. have shown human UBE2M, the E2 for NEDD8, to bind to the Ub-fold domain of UBE1C/UBA3 (Huang, 2005). It would be interesting to see if a novel Ub-binding motif in this E2 with a more general role exists.

There are several type II Ub-like proteins, whose Ub-like domain exhibit significantly more similarity to Ubiquitin than the Ub-fold domains of several E1s do. In some of these type II proteins, the Ub-like domain associates with the proteasome, e.g. in Rad23 and Ubp6 (Borodovsky, 2001, Walters et al., 2002). Taken together, these observations and the results of Huang et al. suggest that the Ub-like domain and the more distantly related Ub-fold domain might generally mediate associations within the UPS as discussed in more detail by Upadhya et al. (Upadhya, 2003).

4.2.2 Subfamilies of the RING superfamily

Originally, the E3 class of RING finger proteins was subdivided into either RING-H2 or RING-HC according to their cysteine/histidine pattern, which is ‘C3H2C3’ or ‘C3HCC3’, respectively. These types of RING finger proteins are widely detectable by available RING finger profiles provided by the Pfam and PROSITE database. Throughout this work, these proteins are referred to as ‘classic’ RING finger proteins to distinguish them from the non-

classic, more divergent ones. In chapter 3.4, the connections of some families of non-classic RING finger proteins to the classic RING finger family have been analyzed and will be discussed now.

The Parkin triad family has been classified as non-classic RING finger family here. Proteins of this family contain three consecutive copies of complex zinc fingers ('Parkin fingers') with weak similarity to the classic RING finger. Moreover, classic RING finger proteins harbour just one RING finger copy. Structurally, at least the third Parkin finger in HHARI/Ariadne is clearly distinct from the classic RING finger and coordinates only one zinc ion, as shown by the crystal structure published by Capili et al. (Capili, 2004).

It is likely that the other Parkin fingers have a similar structure, although this suggestion still requires experimental verification. The three Parkin fingers within a triad are typically quite different and exhibit distinct degrees of divergence from the classic RING fingers. While the third Parkin finger often follows the 'C3HC4' pattern, e.g. Parkin (see Figure 3-5), the remaining two Parkin fingers differ significantly from the classic RING finger. For example, only the first Parkin finger of RNF19 follows the cysteine/histidine pattern of classic RING fingers, while the neighbouring Parkin fingers lack one of the potential zinc coordinating residues or have additional residues inserted between one of the dyads.

A detailed discussion on each of the predicted zinc coordinating residues in all Parkin triad proteins is beyond the scope of this work. As shown in the example PHD fingers in chapter 3.4.4, the consideration of single positions is not an appropriate method for deriving some general insights on structure or function, at least in the case of complex zinc fingers. Another example for a misleading result due to the mere examination of cysteine/histidine patterns is the second Parkin finger, which is often described as being unrelated to the flanking Parkin fingers (Marin et al., 2004). A profile-based analysis of the non-coordinating residues in addition to the coordinating residues rather suggests a relationship to the other Parkin fingers and even to classic RING fingers. Thus, the opinion that the second Parkin finger is different from the other two Parkin fingers should be revised.

Besides the Parkin triad family, the PIAS-type RING finger family, the U-Box family as well as several other proteins collectively referred to as 'degenerate' RING finger have been dealt with in chapter 3.4. At least one member of each of these groups is active as an E3 ligating either Ub or SUMO (Regelmann et al., 2003) (Hatakeyama et al., 2001, Johnson et al., 2001, Marin et al., 2004). These proteins certainly all belong to the RING finger superfamily, but have a different degree of divergence from the classic RING finger. When comparing U-Box proteins with the classic RING finger, the shortcoming of classifying and comparing RING finger

proteins by cysteine spacing alone becomes obvious. U-Box proteins do not contain any potential zinc coordinating residues at positions homologous to zinc coordinating cysteines or histidines in classic RING fingers (see Figure 3-5). Therefore, cysteine spacing is an inappropriate means for quantifying the similarity or dissimilarity of members of the RING finger superfamily. The profile-based approach chosen here relies on similarity scores calculated from all positions within the RING finger and seems to be a better way to estimate the evolutionary connections between the members of the RING finger superfamily. In addition, the profile-based approach provides the opportunity to find more distant members of the RING-finger superfamily in the databases. For a possible function of these proteins as E3s, they have to recruit E2 enzymes. In established RING-type E3s, this recruitment is mediated by the RING finger itself, which forms an interface that binds to the E2. This interface does not contain the coordinated zinc ions in the classic RING fingers, but other residues. The coordinated zinc ions just stabilize the architecture of the E2 binding interface. The non-classic RING finger proteins with an established E3 function rely on other mechanisms than coordinating two zinc ions to stabilize this interface and possibly, novel members of this superfamily will exhibit such mechanisms as well.

4.2.3 Extension of the UBA family

The evolutionary relationship between the UBA- and the CUE-domain has been described in chapter 3.6.1.2. Besides the CUE-domain family, additional UBA-related families could be identified (see Table 3-28). Unfortunately, for the most interesting of these subfamilies (TtrapNT and AriNT), there is no structural information is available and experimental data on the binding capabilities is also missing. When combining both statistically significant sequence similarity and analogous secondary structure predictions, a common 3D structure for the UBA domain and related domains appears likely. In a recently reported crystal structure of an archeaebacterial NACα homologue, the NACαCT-domain clearly adopts a UBA-like fold (Spreter, 2005). In addition, for another UBA-related domain, the TapCT domain, a crystal structure is available, which exhibits a good similarity to the classic UBA domain (Grant, 2003). However, the TapCT domain probably does not bind to Ub but rather to the repetitive Phe-Gly motif found in several nucleoporins (Grant et al., 2003). The binding site of the TapCT domain involved in Phe-Gly motif binding is located at a completely different surface site than the residues found to interact with Ub in the classic UBA domain. This example shows that a proper function in Ub-binding cannot be predicted for every UBA-related subfamily. Probably, the

high divergence of some UBA-like families is associated with the development of completely new binding properties, as observed for the TapCT family (Grant et al., 2003).

4.2.4 A functional role for the USP zinc-finger

The zinc finger that has been found to be inserted in the catalytic domain of USP-type deubiquitylating enzymes (see chapter 3.5.2.4) may give an attractive explanation for unpublished experimental results of Hetfeld et al., who have observed an inhibition of USP15 mediated tetra-Ub cleavage by treatment with OPT, a metallo-protease inhibitor (Hetfeld et al., personal communication). Based on results of this work, a model is likely, in which OPT sequesters the zinc ions needed for the proper conformation of the potential zinc finger. As a consequence, the Ub-binding pocket becomes destabilized. Probably, the zinc removal by OPT does not interfere with poly-Ub cleavage, but maybe with poly-Ub binding. This model would satisfy both the prediction of USP15 as cysteine protease and the prediction of a zinc finger in USP15. Site-directed mutagenesis of the proposed cysteine dyads causes a complete loss of poly-Ub cleavage activity in USP15, similar to the effect of OPT, while a linear Ub-GFP-fusion protein was still cleaved (Hetfeld et al., 2005). This observation points to metal coordination by the cysteines and to an essential but indirect function of the potential zinc finger in poly-Ub cleavage.

4.2.5 Prediction of a common structural scaffold for proteasome lid, COP9-signalosome and eIF3 complexes

4.2.5.1 The bipartite structure of the PCI domain

In the original discovery note, the PCI domain has been defined as a homology domain found within multiple proteins that are otherwise unrelated (Hofmann et al., 1998). The results presented in chapter 3.7.3 suggest that this view should be revised. The sequence regions detected as PCI domains by bioinformatical methods seem to consist of two structurally distinct domains. The C-terminal portion, which in eIF3k is referred to as the WH-domain, is much better conserved in sequence than the N-terminal portion, and the C-terminal boundary of the PCI homology domain is relatively well defined by a notable loss of sequence conservation. By contrast, the N-terminal boundary of the homology domain has always been ill-defined, as the overall sequence conservation in this region is low and different families of PCI proteins appear to lose their similarity at different positions. As a consequence, different domain databases and

their associated web-servers detect PCI-domains (or the synonymous 'PINT' domains) of varying length in the order PROSITE > Pfam > SMART.

Using the eIF3k-derived structural model, most of these observations can be readily explained (see Figure 3-26). The C-terminal PCI/PINT boundary, which is agreed on by all domain databases, corresponds to the C-terminal boundary of the structural WH-like domain. The N-terminal boundary of the PINT domain, as described in the SMART database, essentially corresponds to the N-terminus of the WH-like domain. The PCI domain of the Pfam database corresponds to the WH-portion plus a single α-helical hairpin repeat. Finally, the PCI domain as described in the PROSITE database covers the WH-portion and all three helical hairpin repeats found in the eIF3k structure. Of the three representations, the PINT domain of the SMART database is structurally most correct, as it describes a true autonomously folding domain. The observation that some PCI families lose their sequence conservation at different N-terminal positions can be explained by assuming a variable number of helical repeat motifs for those proteins. As an extreme example, only the WH-like region could be detected in eIF3e by profile searches done here, and the secondary structure prediction for the eIF3e family suggests a β-structure instead of the usual helical-hairpin repeats upstream of the WH region. This finding can be taken as a further hint for structural and functional independence of the N- and C-terminal sub-regions of the PCI homology domain.

4.2.5.2 The nature of the N-terminal helical repeat extension

The finding of TPR-like repeats preceding many PCI domains as described in chapter 3.7.3.2, combined with the helical repeat structure of the N-terminal portion of the PCI domain itself, leads to the interesting question if these repeats are of the same type and may form a continuous solenoid structure. The authors of the eIF3k crystal structure propose a structural relationship between the eIF3k N-terminus and the HEAT motif based on superposition calculations with DALI (Holm et al., 1997). By contrast, sequence-based analysis methods done here rather point to an evolutionary relationship to the TPR motif, both for the region preceding the PCI domain and for the first helical hairpin of the PCI domain itself. A related finding was reported for Rpn3 and Csn12 elsewhere, where a homology domain termed "PAM" (PCI-associated module) with TPR-like properties has been proposed (Ciccarelli, 2003). The findings at hand suggest that the PAM-domain is a special case of a more widespread preference of the WH-portion of the PCI domain to be preceded by TPR-like repeats. In addition, both our results

and those of Ciccarelli et al. argue in favour of a continuity between the N-terminal repeats and those found within the PCI domain.

Due to the borderline sequence similarity between the classical TPR motif and the distinct helical hairpins of the PCI proteins, a completely novel type of bi-helical repeats distinct from TPR and HEAT/Armadillo or some kind of intermediary form cannot be ruled out. Structurally, HEAT and TPR repeats are relatively similar and both tend to form superhelical solenoid structures (Kajava, 2002). Without assuming a particular repeat family, I have attempted a rough estimation of what a typical PCI component of the lid or the CSN complex might look like. Figure 4-1 shows schematically a PCI protein with a WH-like domain at the C-terminus (green), preceded by three helical-repeats assumed to lie within the PCI boundaries according to PROSITE (dark blue), which are in turn preceded by three additional helical repeats that represent the TPR-related N-terminal extension (light blue). As I do not assume a particular repeat family with a well-known radius of solenoid curvature, I used the values derived from the first two helical hairpins of the eIF3 structure instead. It should be stressed that the model of Figure 4-1 with its 'boomerang'-shaped architecture can only give a very coarse approximation of the real situation. Both the solenoid curvature and the exact number of N-terminal repeat extensions are rough estimations. Nevertheless, the model appears to be roughly compatible with the electron density maps of the lid and CSN complexes (Kapelari et al., 2000).

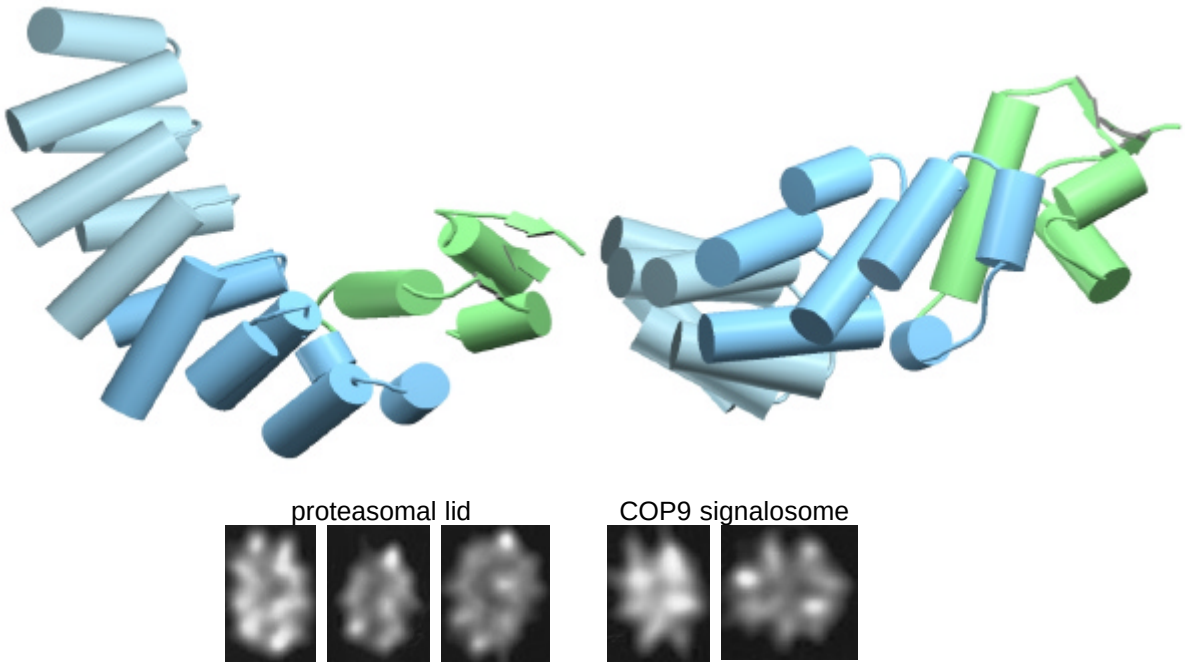


Figure 4-1 Model of a PCI protein with three additional helical hairpins upstream of the PCI domain. Within the PCI domains, only regions that can be modelled on the eIF3k template are shown. The N-terminal extension is shown in light blue, the other colours are as in A. Regions belonging to the WH subdomain are shown in green, while conserved structure elements of the helical hairpin regions are shown in dark blue. Other regions (extreme N- and C- termini, connecting helices between hairpins, unstructured regions) are shown in grey. In the lower panel,

electron microscopic images of the proteasomal lid and the COP9 signalosome, CSN, are depicted (Kapelari et al., 2000).

4.2.5.3 A structural scaffold for three multi-protein complexes

PCI proteins constitute the main components of the proteasome lid and the CSN complex and also form the structural core of the translation initiation factor eIF3. So far, no catalytic activity has been described for PCI proteins. Given the lack of invariant polar residues, such a role appears unlikely. The role of the PCI domains is most likely that of a scaffold for the other complex subunits and other binding partners. There are at least three distinct structural roles that PCI proteins have to fulfil: i) maintaining the integrity of the complex by binding to other PCI proteins, ii) attaching the MPN-subunits to the complex, and iii) binding to other partners such as the base-complex in the case of the proteasome lid or the RNA-binding subunits of the eIF3 complex.

The assignment of these functionalities to the different regions of the PCI proteins, and equally important, the source for the subunit interaction specificity or promiscuity have been subject to several experimental studies, some of them published while others have been presented at a recent meeting on PCI complexes (Chang, 2004). The PCI model presented here will be certainly useful, both for the interpretation of the experimental results, and for the design of new experiments e.g. those based on domain truncations or domain swaps. According to the analysis carried out here, in some proteins the PCI domain is restricted to a C-terminal WH-like part. As these proteins are also components of PCI-complexes, a role of the WH domain in PCI:PCI domain interaction is very likely. On the other hand, TPR-repeats in general form versatile protein-interaction surfaces and the same is expected to be true for the TPR-like repeats found in the PCI proteins (D'Andrea et al., 2003).

Tsuge et al. analysed truncated forms of human Csn1 for its interaction with other PCI subunits of the CSN complex (Tsuge et al., 2001). A construct containing residues 197-500 (corresponding to the entire PCI region and some C-terminal material) was able to bind to Csn2, Csn3 and Csn4. Another construct starting at position 340, and thus lacking the helical-repeat region, no longer bound to Csn2 and Csn4 but maintained binding to Csn3. By contrast, a construct 197-307 that lacks the WH-like region was only able to bind to Csn4. These experiments suggest that both the WH portion and the helical-repeat part of the PCI proteins have a role in PCI:PCI interactions, although they seem to interact with different subunits of the complex. The importance of both subdomains is confirmed by a recent study of Isono et al., who analyse multiple point mutations in the lid subunit Rpn7 (Isono, 2004). In their hands, both

mutations in the N-terminal helical repeat part of Rpn7 and mutations within the WH-like region are able to abrogate binding to Rpn3, another PCI subunit of the proteasome lid. So far, no information is available on the PCI regions involved in binding to the MPN subunits.

In summary, I believe the PCI domain could play a role as a universal binding domain supporting intra-complex interactions as well as recruitment of additional ligands. The model presented here is a first step to the understanding of the supramolecular architecture of three important complexes and certainly will facilitate the interpretation of further experimental results. Nevertheless, a full understanding of the interaction mode between PCI- and MPN-domain proteins will certainly require experimentally determined high-resolution structures of the components - or ideally, that of an intact complex.

4.2.6 Ubr1 might use its ClpS domain for proteasome binding

As described in chapter 3.7.6.6, both ClpS and Ubr1 share a common homology domain ('ClpS domain') and bind to AAA-ATPase subunits of ClpAP and the proteasome, respectively, two analogous chambered proteases (Dougan et al., 2002, Xie et al., 2000). Both Ubr1 and ClpAP function in the N-end-rule pathway (Varshavsky, 1996, Xie et al., 2000). These findings suggest that the ClpS domain in Ubr1 might mediate the binding of Ubr1 to the Rpt6 and Rpt1 subunits of the proteasome. However, the binding mode of the ClpS-ClpA interaction seems not conserved in Ubr1-proteasome association, as neither the site of ClpS needed for binding to ClpA is conserved in the ClpS domain of Ubr1 nor is the ClpA binding site for ClpS recognizable in any proteasomal subunit. Nonetheless, by sequence analysis alone, a similar binding mode for Ubr1-proteasome association should not be ruled out.

The ClpS domain in Ubr1 homologues probably adopts the same conical fold as ClpS from *E. coli*. As ClpS is obviously a monolithic protein, the ClpS domain in Ubr1 should fold independently from the remaining residues of Ubr1. Therefore, the expression of a single Ubr1-ClpS domain should be possible, which would allow binding studies with this domain and proteasomal subunits. These experiments may help answering the question about the role of the ClpS domain in the binding of Ubr1 to the proteasome.

4.3 Comparing the yeast and human UPS

The protein families compiled in chapter 3 provide useful information for comparing the yeast and human UPS on a more general level than just determining orthologues. For that purpose, protein families were grouped according to their main function within the UPS (Figure 4-2) and data on orthology assignments from chapter 3 was used to define the fraction of each group without orthologue in yeast or human ('orphans'), respectively.

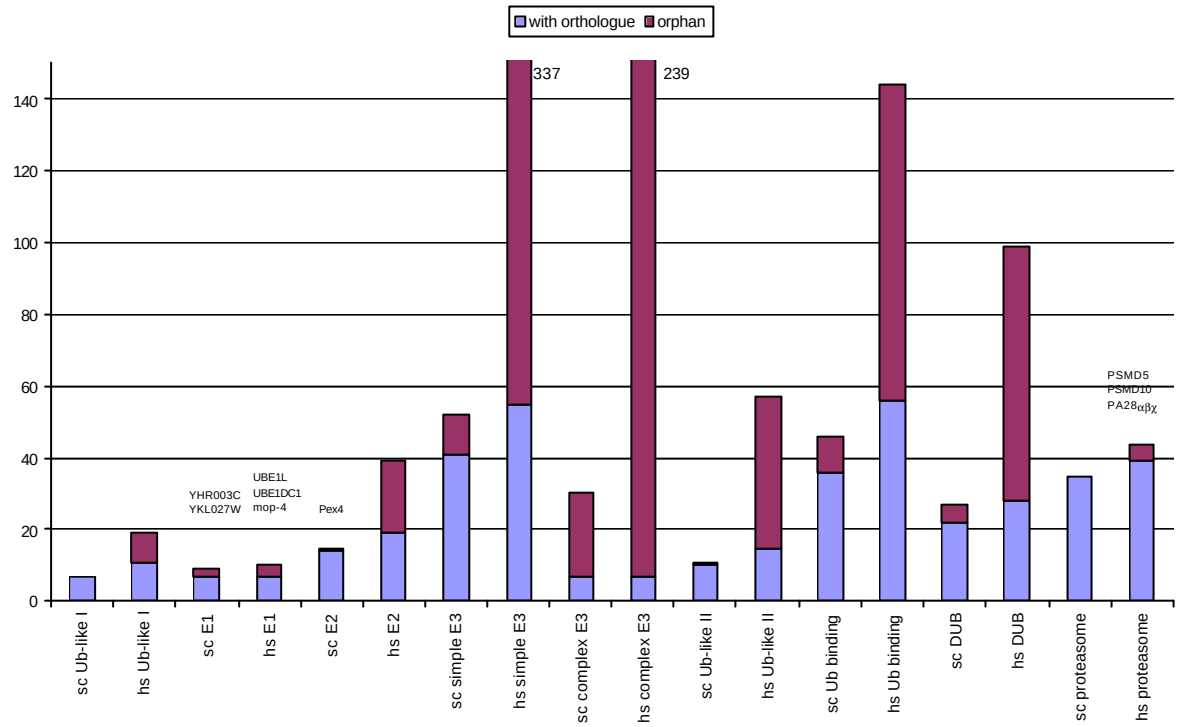


Figure 4-2 Distribution of protein class sizes for all major classes of the UPS. The protein families contributing to each protein class are described in the text. Protein names above some bars denote the orphans in the corresponding class. The y-axis indicates the number of group members. 'sc' = *S. cerevisiae*, 'hs' = *H. sapiens.*

4.3.1 Ub family, E1 and E2 enzymes

Number of orthologues may differ within a group

The yeast genome encodes seven type I Ub-like modifiers and each of them could be assigned to at least one orthologue by means of sequence analysis. For the human orthologues of the yeast modifiers Smt3 and Atg12, gene duplications have taken place, leading to 1:2 and 1:4 orthologous relationships, respectively. Species-specific gene duplications, as found for these two proteins, are a general observation in all other protein families examined. As a consequence, the number of proteins with orthologues in a given group of one species may differ from the corresponding number in the second species as shown in Figure 4-2.

More modifiers than E1s in human

In the human genome, 19 Ub-like modifiers could be detected, some of them still with a status as putative modifier. Of these 19 genes, eight were classified as orphans. Interestingly, the number of human E1 genes (10) found by sequence comparison turned out to be much smaller than the number of Ub-like modifiers. This finding may indicate redundant activation mechanisms or so-far unexplored classes of E1 proteins. At least for Atg8 and Atg12, a common E1 exist, which is Atg7 (Ohsumi, 2001).

Yeast-specific E1s without modifiers?

The human genome contains three E1 genes that are not present in yeast: UBE1L, UBE1DC1/Uba5 and mop-4 (see Table 3-1). UBE1L and UBE1DC1/Uba5 activate human-specific ISG15 and Ufm1, respectively, while the substrate of the third activator is unknown. Interestingly, mop-4 exhibits a domain structure identical to UBE1L. It is conceivable that mop-4 activates the second di-ubiquitin-modifier in human, Fat10, which is absent in yeast and has so far no assigned E1.

There are two yeast-specific E1 proteins, which are so far uncharacterized ORFs (see chapter 3.2). As all yeast modifiers except for Hub1 have known E1s (see Table 3-1) and these two ORFs have no known substrate so far, they might be responsible for Hub1 activation. However, mutagenesis of these two ORFs shows no effect on post-translational modification by Hub1 (G. Dittmar, personal communication). Besides, a role of Hub1 as a true covalent modifier is controversial (Luders et al., 2003, Yashiroda et al., 2004). It is not clear whether these two orphan activators act on known yeast Ub-like modifiers or whether they fulfil a totally unrelated function.

Proportions for E2s, E1s and modifiers do not reflect proteome ratio

The human genome encodes more E2 proteins than the yeast genome, although this ratio is by far smaller (~2.5:1) than would be suggested by the estimated proteome sizes from human and yeast (estimated to be between 6:1 and 7:1, here named ‘proteome ratio’). Thus, the expansion of human E2 does not appear extraordinarily here. The same is valid for human Ub-like modifiers and E1s. However, the situation for the E3 enzymes is clearly different as shown in the following chapters.

4.3.2 Simple and complex E3 enzymes

Proportions for yeast and human E3s correspond to proteome sizes

The E3s are subdivided into two groups, one comprising RING/HECT/A20-zinc-finger E3s ('simple E3s'), the other one containing the substrate binding subunits of Cullin/RING-based E3s ('complex E3s'). The main contributors to the latter group are F-Box, BTB and SOCS proteins as described in chapter 3.4.5. Unlike the groups of Ub-like modifiers, E1s and E2s, the proportions of human and yeast simple E3s as well as of human and yeast complex E3s correspond approximately to the proteome ratio (see Figure 4-2). Notably, both E3 groups are remarkably large in size as compared to 'upstream' components of the ubiquitylation cascade, E1s and E2s.

Not all yeast E3 families have been expanded since the splitting of the metazoan and the fungal lineages

A closer inspection of the protein families in the group of complex E3s allows further insights into the evolutionary events that led to the large total number observed here. Yeast and human are representatives of the fungal and metazoan lineage, respectively. The high percentage of yeast genes with human orthologues in the group of simple E3s might indicate an already expanded group of simple E3s in the common ancestor. Similarly, this ancestor likely encoded members of all families within the group of complex E3s, which are the F-Box, the SOCS and the BTB family (see Figure 4-3).

There are two hints that each complex E3 family contained only few members in the common ancestor. First, the small amount of BTB and SOCS proteins in yeast might indicate that at least these two families were of the same size in the common ancestor. Secondly, the number of assignable orthologues is notably small for all the F-Box, the SOCS and the BTB families (see Table 3-17), suggesting that only representatives of these genes might have been present in the common ancestor. However, this evolutionary scenario is only true if the large-scale loss of F-Box, SOCS and BTB members in the fungal lineage can be ruled out.

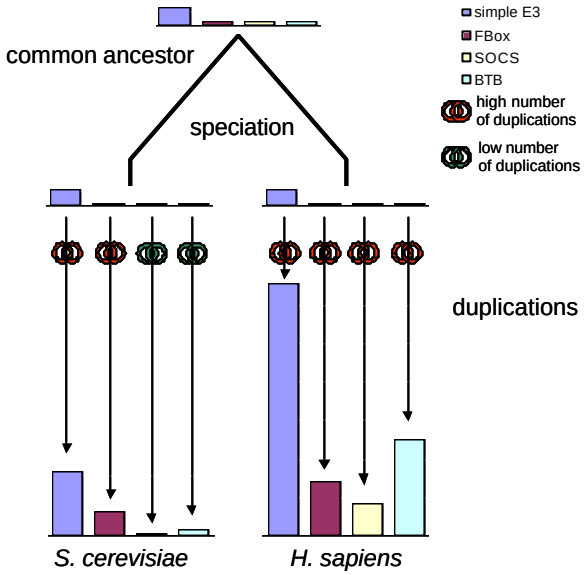


Figure 4-3 Evolutionary model to explain the different portion of orphans in the group of 'simple E3s' and 'complex E3'. The common ancestor had the 'simple E3s' already expanded, while expansion of the families that contribute to the 'complex E3' group took place not until splitting of the fungal and metazoic lineage (here indicated as 'speciation').

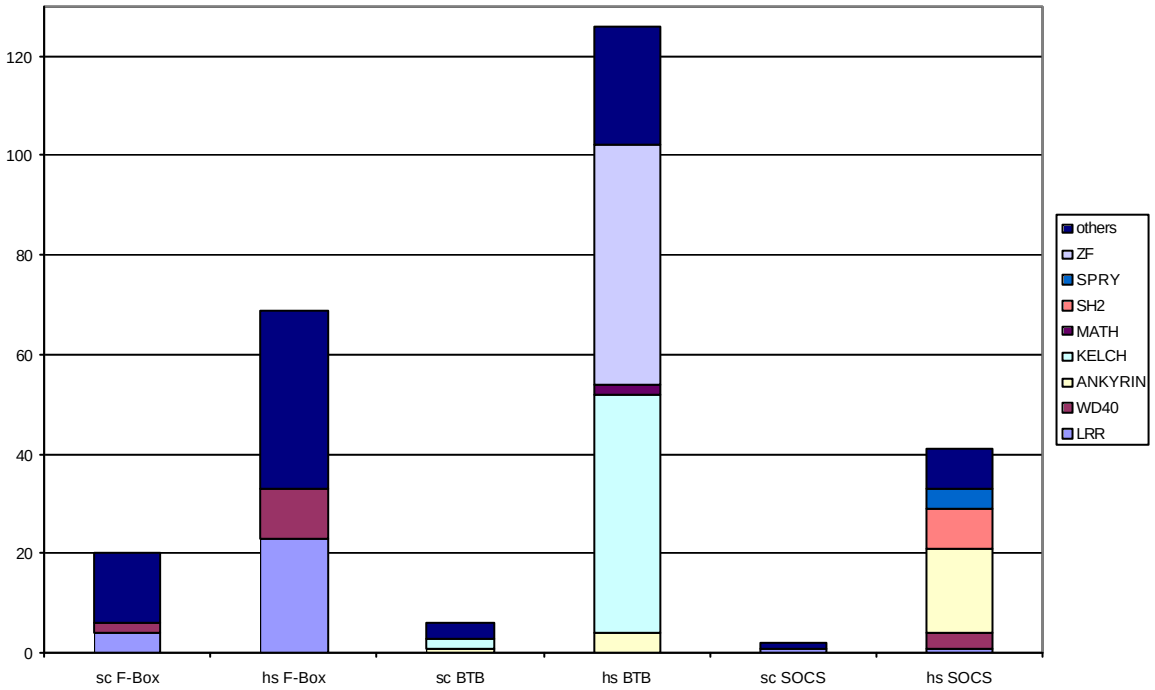


Figure 4-4 Distribution of substrate-binding subunits of cullin/RING-based E3s ('complex E3s'). The subunits were grouped depending on the domain (F-Box, BTB or SOCS) that mediates contact with the cullin/RING-scaffold. Groups were further subdivided based on the known or putative protein-interaction domain that binds to the substrate (see legend). The y-axis indicates the number of members in each group. 'sc' = *S. cerevisiae*, 'hs' = *H. sapiens.*

In yeast, the F-Box family accounts for 20 out of the 28 complex E3s, while SOCS and BTB proteins are only found in small amounts (see Figure 4-4). Both yeast SOCS proteins,

Rad7 and Ela1, have human orthologues. According to the evolutionary model described above, this finding suggests that in budding yeast or the fungal lineage in general no gene duplication has taken place for the SOCS family. Of course, high evolutionary rates of duplicated SOCS genes in yeast might conceal their membership of the SOCS family and thus the real number of SOCS genes might be larger than two genes reported here. Nonetheless, gene duplications in the yeast SOCS and BTB families seem to be very rare in comparison to human.

Orthologous pairs of complex E3 contain proteins with roles in central processes

The number of genes in the complex E3 group is relatively high. In yeast, only the F-Box family has been expanded, while in human, genes of all three classes underwent manifold duplications. The duplications of these genes have probably allowed to make new substrates accessible to Ub-dependent pathways.

After a duplication event, often one gene copy keeps the original function while the other one undergoes neo-functionalization and diverges to a greater extent. This more general observation plays a role in interpreting the finding that only few orthology assignments could be set up for all three subfamilies of the complex E3 group in this work. Most of the complex E3s with an orthologue carry out the same essential functions in yeast and human. For example, Ela1 and its human orthologues have a role in transcription elongation (see Table 3-17). In addition, yeast Cdc4 and human FBXW7 are both involved in cyclin degradation and are therefore crucial for cell cycle progression. Like the previous examples, most of the remaining substrate-binding subunits with orthologues perform central cellular functions (e.g. DNA damage repair or translation elongation). It is likely that all three families trace back to essential genes already important in the common ancestor and that they have been kept at least as one copy with the original function.

Complex E3 family expanded more than simple E3 family

The portion of species-specific members of the complex and simple E3 families is indicative of the point in evolution when the expansion of the groups began. The RING and HECT families contain a high proportion of genes with orthologues in the other species and have been expanded to a substantial degree before splitting of the metazoan and fungal lineage (see Figure 4-3). Otherwise, family expansion of all F-Box, SOCS and BTB families took place after the splitting. Since the splitting of the two lineages, expansion of the simple E3s has proceeded at the same rate in both yeast and human, because the proportion of yeast simple E3s to human simple E3s follows the proteome ratio. The small number of orthology assignments suggests that the basis of SOCS, F-Box and BTB family expansion were only few genes. As can

be seen from the high number of members in the complex E3 group, the expansion of this group seems to have proceeded more vigorously than the expansion of the group of simple E3s.

Complex E3 group evolved differently in human than in yeast

While the ratio of yeast simple E3s and yeast complex E3s is ~1.9, the analogous ratio in human is just ~1.4. This observation may be interpreted as a more vigorous expansion of complex E3s in human compared to yeast, given that the group of simple E3s evolved at the same rate in both species. As will be discussed below, the families within the group of complex E3s have contributed differently to expansion (see Figure 4-4). It is still not clear, why evolution has placed more emphasis on families of substrate-binding subunits of complex E3s than on families from the evolutionary older group of simple E3s. An explanation may be found in the observation that F-Box, BTB and SOCS proteins quite often employ substrate-binding domains made from repetitive elements, such as KELCH domains, WD40 regions or leucine-rich repeats (LRRs) (see Figure 4-4). These repetitive domains probably acquire completely new binding properties more easily than the compact substrate-binding domains. The repeats in the proteins are encoded by corresponding repetitive DNA segments, which are prone to unequal crossing-over leading to additional repeats. While new repeats might be inserted this way, the original substrate-binding site has a high chance to remain untouched. As a consequence, these types of mutations should have been manifested more easily in the genome and novel repeats could evolve new binding properties later on. On the other hand, the gain of new-binding properties in RING or HECT proteins is more likely based on single point mutations affecting the non-repetitive substrate-binding domain.

Interestingly, repetitive bihelical regions are also found in the PCI subunits of the proteasomal lid and the CSN complex (see chapter 3.7.3.4). Both complexes contain binding sites for a variety of other proteins, exceeding even the number of the subunits in each complex. It is possibly that evolutionary events, as proposed for the subunits of complex E3s described above, which use repetitive regions for substrate-binding, have also led to distinctly elongated repetitive regions in PCI subunits as described in chapter 3.7.3.2.

Emphasis on F-Box-family evolution in yeast compared to human

An interesting observation can be made, when comparing the expansion of F-Box proteins in yeast to the expansion of SOCS and BTB proteins in human. It appears that after splitting of the fungal and metazoan lineages, yeast has expanded its repertoire of F-Box proteins more vigorously than expected. In human, too, a certain expansion of the F-Box family has taken place, but compared to the proteome size, it is not as pronounced as in yeast. It is interesting to observe that in other organisms including the plant *A. thaliana* and the nematode

C. elegans the F-Box protein repertoires have been expanded in particular (data shown). In human, the sets of SOCS and BTB proteins have been dramatically expanded, a development that is not observed at all in yeast. It is conceivable that certain biological functions are carried out by an F-Box protein in yeast, while in human a SOCS or BTB protein is used for the same purpose.

Novel roles for ancient substrate binding subunits?

The vast expansion of protein families from the E3 group might have coincided with the establishment of completely new functions outside the UPS or even outside Ub-dependent processes. In this respect, the multiplicity of combinations between the RING/cullin/SKP1-scaffold binding domain and other domains is remarkably, especially for BTB and SOCS proteins. For example, a significant number of SOCS proteins harbour an SH2 domain known to bind phosphotyrosine-containing proteins, which are often involved in signalling cascades. Moreover, a huge portion of BTB proteins have acquired repetitive regions composed of C2H2-type zinc-fingers, which originally have been thought to bind to nucleic acids (Evans, 1988). So far, none of the BTB/C2H2 type proteins has been reported to be an active subunit of a complex E3 or to bind to nucleic acids. Therefore, a role for this subfamily of BTB proteins in Ub-dependent processes remains elusive.

No yeast orphans for HECT and atypical RING proteins

A subclassification of the RING superfamily into classic RING finger, Parkin triad, PIAS finger, U-Box and degenerated RING finger shows that yeast-specific genes are only observed within the group of classic RING fingers (see Figure 4-5). By contrast, human-specific genes are found in all RING subfamilies with the Parkin triad group displaying a pronounced expansion compared to yeast. One explanation for this expansion might be a tissue-specific expression of these genes. In fact, several of the human-specific Parkin triad proteins display an above-average expression in brain (Parkin, RNF144, IBRDC1) or cortex (ARIH2, PARC). These proteins might play a role in processes specific to the nervous system. Indeed, mutations in the Parkin gene are causative for the autosomal juvenile form of Parkinson's disease, PDJ, which is connected with the loss of dopaminergic neurons (Kitada, 1998). Moreover, the Parkin protein is found in Lewy bodies of patients suffering from the sporadic form of Parkinson disease (PD) and Parkin has been assigned an essential role in the generation of these Lewy bodies (Chung, 2001).

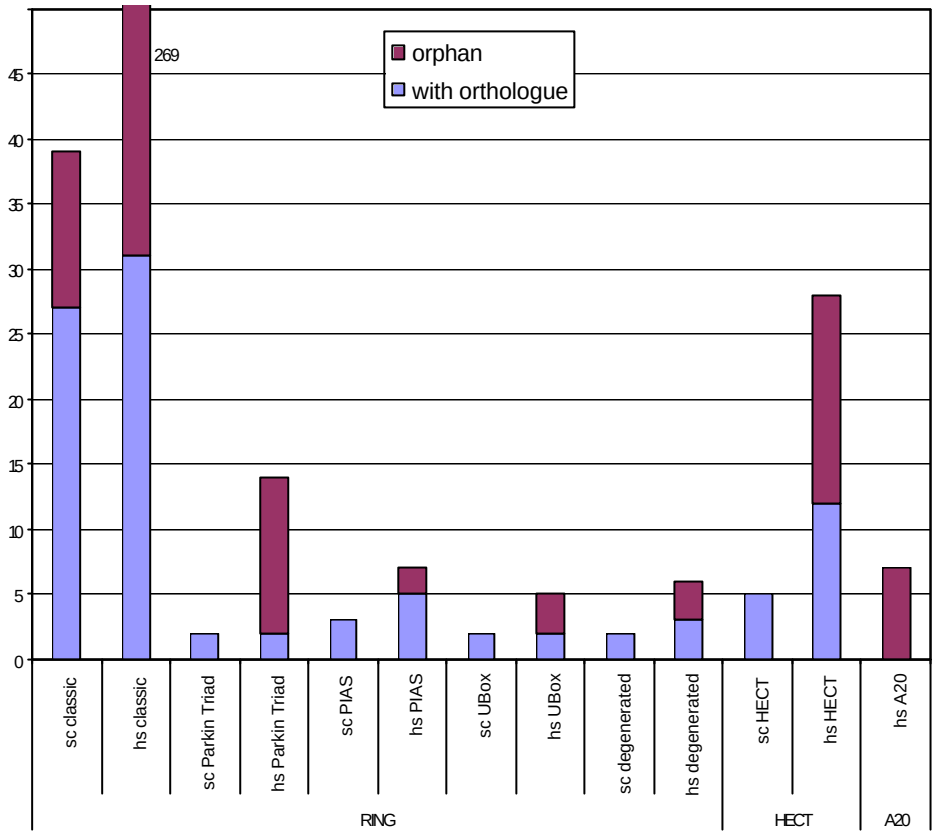


Figure 4-5 Distribution of 'simple E3s'. This group comprises RING-, HECT- and A20-type ligases. The RING-type ligases were further subdivided based on the subfamilies. The y-axis indicates the number of family members. 'sc' = *S. cerevisiae*, 'hs' = *H. sapiens.*

The gene family of A20-zinc fingers is not found in yeast

No A20-zinc fingers could be detected within yeast, while seven clearly paralogous A20-zinc finger proteins were found in human (see Figure 4-5). Cases like this, where yeast lacks a complete family that has a role in the human UPS, are rare. Another example is the Josephin family, which has four human members. It is possible that the common ancestor of the yeast and human lineage had all the extent UPS-relevant families, but specific gene-loss within the yeast lineage resulted in the observation made for A20-zinc-finger proteins and Josephins. Of course, a gain-of-function of families not involved in the common ancestor's UPS could be an explanation as well. A more detailed analysis of other species than yeast and human might give insights on which way was chosen in evolution.

4.3.3 Deubiquitylating enzymes

Within the distinct families of yeast DUBs, only the USP family harbours some yeast-specific genes (Figure 4-6). The members of all other families have human orthologues or are absent in yeast (Josephins). For example, the UCH family has only one member in yeast, Yuh1, while the human genome encodes four UCH proteins. A possible explanation for the human-

specific expansion of the UCH family would be a role in the processing of human-specific Ub-like modifiers. However, processing of Ub precursors in yeast can be performed by many DUBs and none of them, except for Rpn11, is essential for viability (Amerik et al., 2000). Human UCHL5/UCH37 is found associated with the regulatory particle of the proteasome and edits poly-Ub chains conjugated to substrates, a role clearly beyond precursor processing (Lam et al., 1997, Lam et al., 1997). Therefore, the additional human UCH family members do not necessarily correlate with the additional human Ub-like modifiers.

Within the group of DUBs, the OTU family exhibits the most pronounced expansion in the human genome as compared to the yeast genome. The ratio of human to yeast OTU proteins is 7:1, while for USPs, MPNs and ULPs the analogous ratio is just ~3. Nonetheless, the OTU-based ratio still corresponds to the proteome ratio. It is remarkable that the human OTU family expansion is accompanied by new domain topologies. For example, the OTU domain can be found together with Ub-binding domains like the NZF domain or the UBA-related TtrapNT domain (see chapter 3.6). Moreover, in TNFAIP3/A20, as well as two Cezanne paralogues, the OTU domain is combined with the ZF_A20 domain, which at least in TNFAIP3/A20 confers E3 activity (see chapter 3.6.5) (Evans et al., 2004, Wertz et al., 2004).

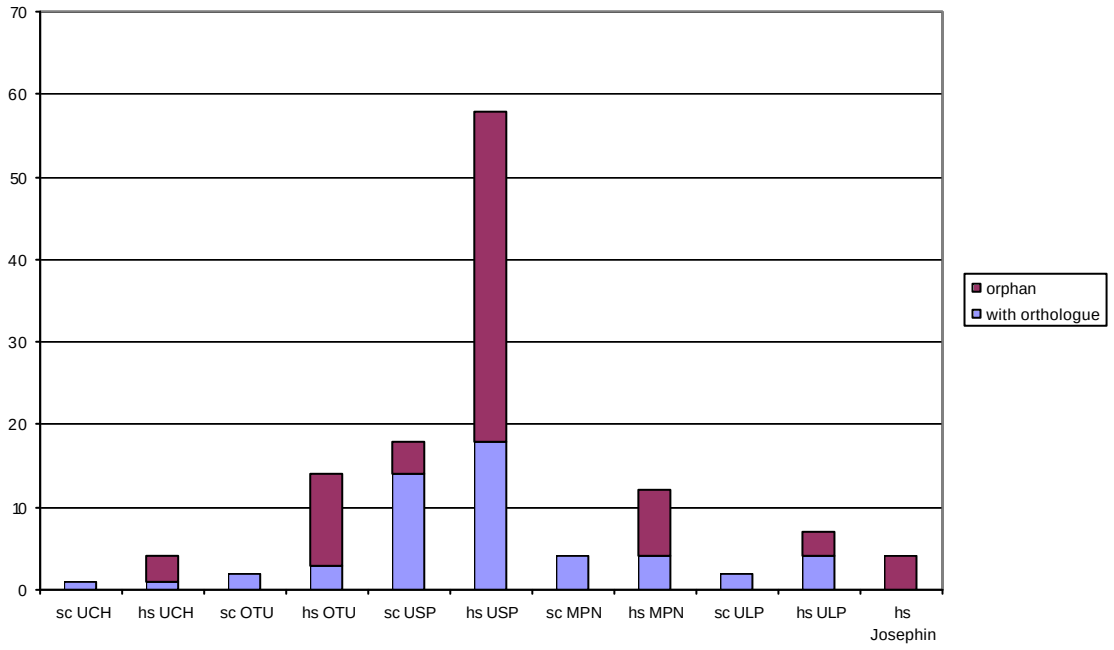


Figure 4-6 Distribution of the known DUB classes including the ULP family. The y-axis indicates the number of family members. 'sc' = *S. cerevisiae*, 'hs' = *H. sapiens*.

4.3.4 Ub-binding proteins

The UBA and the UIM proteins constitute the largest Ub-binding protein families in both yeast and human. In the UBA family, yeast shows nine genes without human orthologues,

while in the UIM family there is only one, Ufo1 (Figure 4-7). The F-Box protein Ufo1, which harbours three UIM copies, is responsible for ubiquitylation of the phosphorylated Ho endonuclease, which has a role in mating type switching and gene-conversion at the mating-type locus (Kaplun, 2003, Kostriken, 1983).

The remaining families do not contain any yeast specific genes and are rather small, except for the human NZF family, which has 19 members compared to two in yeast. As mentioned in chapter 3.6.4, a general role for the NZF domain in Ub-binding is currently not really clear. Thus, the NZF family expansion in the human lineage is not necessarily linked to the evolution of the human UPS.

The difference between the yeast and human PAZ-family might be explained by the expansion of the USP family that contributes most members to the PAZ-family.

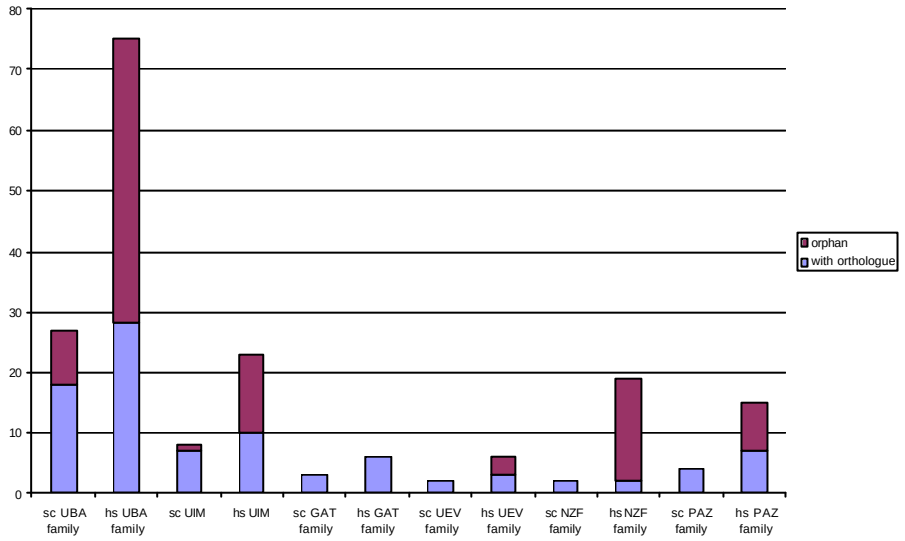


Figure 4-7 Distribution of Ub-binding proteins. The UBA family includes proteins with a UBA-related domain. The y-axis indicates the number of family members. 'sc' = *S. cerevisiae*, 'hs' = *H. sapiens.*

4.3.5 Comparison of PCI complexes regarding their PCI/MPN architecture

'PCI complexes' is a collective term for proteasome lid, the CSN complex and eIF3 as introduced in chapter 3.7.3. While the lid and the CSN complex in human as well as the lid in yeast have an analogous architecture of six PCI-subunits and two MPN-subunits ('6+2'), the yeast CSN complex is quite different fom its human counterpart (see Table 4-1). First, the yeast CSN exhibits just a '4+1' architecture with Csn5 as MPN+ subunit. Second, it contains an unusual PCI subunit, Csn12, with an important role in maintaining the integrity of the complex (Maytal-Kivity et al., 2003). The human orthologue of Csn12 has so far not been reported as a subunit of the human core CSN complex.

Table 4-1 PCI and MPN subunits of CSN and proteasomal lid in human and yeast.

Domain	human lid	yeast lid	human CSN	yeast CSN
PCI	PSMD6	Rpn7	Csn1	Pci8/Csn11
PCI	PSMD11	Rpn6	Csn2	Rri2/Csn10
PCI	PSMD3	Rpn3	Csn3	-
PCI	PSMD12	Rpn5	Csn4	-
PCI	PSMD13	Rpn9	Csn7a/Csn7b	Csn9
PCI	PSMD8	Rpn12	Csn8	-
PCI			Csn12	Csn12
MPN+	Rpn11, PSMD14	Rpn11	Csn5	Rri1/Csn5
MPN	Rpn8, PSMD7	Rpn8	Csn6	-

Of the human PCI complexes, the more distantly related eIF3 complex has only three readily detectable PCI proteins: eIF3a (EIF3S10), eIF3c (EIF3S8) and eIF3e (EIF3S6) (see Table 4-2) (Hofmann et al., 1998). Recent work by Morris-Desbois et al. (Morris-Desbois et al., 2001) has grouped eIF3l (EIFS6IP) with the PCI components of eIF3, and results from this work as described in chapter 3.7.3.3 add eIF3k (EIF3S11) to the ranks of PCI proteins. Besides the PCI subunits, vertebrate eIF3 complexes also contain two MPN proteins: eIF3f (EIF3S1) and eIF3h (EIF3S3). Unlike the situation in the lid and CSN complexes, both MPN subunits of eIF3 have lost their metal-coordinating residues and are most likely catalytically inactive. In addition, yeast and several other unicellular eukaryotes do not seem to have any eIF3-associated MPN proteins.

Table 4-2 PCI and MPN subunits of the eIF3 complex in human and yeast.

Domain	human eIF3	yeast eIF3
PCI	eIF3a, EIF3S10, p170	Rpg1/Tif32
RBD	eIF3b, EIF3S9, p116	Prt1
PCI	eIF3c, EIF3S8, p110	Nip1
	eIF3d, EIF3S7, p66	-
PCI	eIF3e, EIF3S6, p48	-
MPN	eIF3f, EIF3S5, p47	-
RBD	eIF3g, EIF3S4, p44	Tif35
MPN	eIF3h, EIF3S3, p40	-
WD40	eIF3i, EIF3S2, p36	Tif34
	eIF3j, EIF3S1, p35	Hcr1
PCI	eIF3k, EIF3S11, p28	-
PCI	eIF3l, EIF3S6IP, p69	-

Comparing the '5+1' stoichiometry of human eIF3 with the two better-conserved PCI complexes, only one PCI subunit seems to be missing. The sequence analysis efforts in this work have also included other known eIF3 subunits, but no indications for further PCI domains could be obtained (data not shown). Given the high degree of PCI sequence divergence, it cannot be fully excluded that one of the non-PCI/non-MPN subunits (eIF3b, eIF3d, eIF3g,

eIF3i, eIF3j) harbours a cryptic PCI domain that has eluded its detection. On the other hand, it is well conceivable that eIF3 has a deviating subunit composition. In yeast and several other organisms, not only the MPN proteins are missing but also the number of PCI components is reduced, as eIF3e and eIF3l are absent. At present, it is not clear whether the corresponding positions in the complex are left empty or are filled by additional copies of the remaining PCI components. In evolutionary terms, it appears likely that the eIF3 complex is a 'degraded' copy of an ancient lid-like complex, which has lost its MPN+/JAMM mediated catalytic activity and potentially some of its PCI subunits. In yeast, the loss of PCI subunits has progressed more rigorously than in human, as only a '2+0' architecture is recognizable. In turn, by acquiring a group of novel non-PCI/non-MPN subunits, e.g. RNA binding proteins, the eIF3 complex has gained a functionality that is different (and potentially even completely unrelated) to the proteasomal lid and the CSN complex.

4.4 Conclusions and future directions

In this work, proteins harbouring homology domains with relevance to the UPS or other processes involving Ub or Ub-like modifiers were catalogued for yeast and human. As pointed out in chapter 4.1, profile-based means were particularly useful for an exhaustive detection of these protein sets. Additionally, profiles were successful in determining novel homology domains or motifs with importance for processes involving Ub or Ub-like modifiers. All profiles refined or *de novo* defined in this work have the advantage to be usable for rapid screening of other proteomes and novel protein sequences. Therefore, efforts to extend the catalogue of UPS-relevant proteins to other species should be feasible. Broadening the spectrum of evolutionary lineages would certainly give a more comprehensive and detailed view on the processes taken place during UPS evolution.

As the compilation of proteins and their functional classification was determined purely by bioinformatical means, many of the results of this work should be considered as predictions. As a consequence, experimental validation is needed like it has already been carried out for the prediction of ataxin-3 to have a deubiquitylating activity (Burnett et al., 2003, Scheel et al., 2003). In this context, it would be interesting to see if orthologues defined here really execute comparable cellular functions as it has been shown for the yeast orthologue of human antizyme (Palanimurugan et al., 2004). There is a good chance that functional annotation mapped via the homology approach onto orthologues might give reliable predictions in general.

Some of the future work should deal with uncovering the roles of additional, not directly UPS-relevant domains in proteins mentioned in this work, especially in E3 ligases and in proteins with Ub-like or Ub-binding domain(s). Besides a better subclassification and therefore clarity of these protein classes, this type of examination would allow insights to how Ub-related processes became integrated into signalling, endocytosis, etc. and how these processes contribute to the complexity and diversity of multicellular organisms.

Short motifs such as the SUMO interacting motif appear underrepresented as compared to larger homology domains described in chapter 3. In general, the number of known short motifs is small and only few are well annotated (Puntervoll, 2003). One reason for this is that the shortness of these motifs places limitations on the specificity of the profile searches. Given a certain short motif, the chance to detect an analogous motif that has evolved accidentally is negatively correlated with the motif length. Nonetheless, at least the generalized profile technique in combination with additional criteria as outlined in chapter 3.6.7.3 is an adequate means to cope with short motifs yet to be discovered.

Establishing a catalogue of UPS-relevant proteins may be a starting point for high-throughput approaches with focus on a pre-defined set of genes or proteins. In this respect, gene expression profiling based on microarrays will be of particular interest as this type of analysis allows the discovery of genetic interactions (pathways) and genes acting as groups as well as the functional elucidation of individual genes. Besides providing genes of interest for designing microarrays, functional annotation inferred from homology domain analysis or from orthologues will be useful for interpretation of microarray experiments (pathway analysis) and other large-scale data. Another application of the catalogue of UPS-relevant proteins may be a directed RNAi analysis as it has already been carried out in *C. elegans* (Jones, 2001).

5 References

Ahmad KF, Engel CK, Prive GG: **Crystal structure of the BTB domain from PLZF**. *Proc Natl Acad Sci U S A* 1998, **95**(21):12123-12128.

Alam SL, Sun J, Payne M, Welch BD, Blake BK, Davis DR, Meyer HH, Emr SD, Sundquist WI: **Ubiquitin interactions of NZF zinc fingers**. *EMBO J* 2004, **23**(7):1411-1421.

Alberti S, Demand J, Esser C, Emmerich N, Schild H, Hohfeld J: **Ubiquitylation of BAG-1 suggests a novel regulatory mechanism during the sorting of chaperone substrates to the proteasome**. *J Biol Chem* 2002, **277**(48):45920-45927.

Albrecht M, Hoffmann D, Evert BO, Schmitt I, Wullner U, Lengauer T: **Structural modeling of ataxin-3 reveals distant homology to adaptins**. *Proteins* 2003, **50**(2):355-370.

Altschul SF, Carroll RJ, Lipman DJ: **Weights for data related by a tree**. *J Mol Biol* 1989, **207**(4):647-653.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

Amerik A, Swaminathan S, Krantz BA, Wilkinson KD, Hochstrasser M: **In vivo disassembly of free polyubiquitin chains by yeast Ubp14 modulates rates of protein degradation by the proteasome**. *Embo J* 1997, **16**(16):4826-4838.

Amerik AY, Li SJ, Hochstrasser M: **Analysis of the deubiquitinating enzymes of the yeast Saccharomyces cerevisiae**. *Biol Chem* 2000, **381**(9-10):981-992.

Anantharaman V, Aravind L: **Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes**. *Genome Biol* 2003, **4**(2):R11.

Andrews EA, Palecek J, Sergeant J, Taylor E, Lehmann AR, Watts FZ: **Nse2, a component of the Smc5-6 complex, is a SUMO ligase required for the response to DNA damage**. *Mol Cell Biol* 2005, **25**(1):185-196.

Aravind L, Iyer LM, Koonin EV: **Scores of RINGS but no PHDs in ubiquitin signaling**. *Cell Cycle* 2003, **2**(2):123-126.

Aravind L, Koonin EV: **The U box is a modified RING finger - a common domain in ubiquitination**. *Curr Biol* 2000, **10**(4):R132-134.

Aravind L, Ponting CP: **Homologues of 26S proteasome subunits are regulators of transcription and translation**. *Protein Sci* 1998, **7**(5):1250-1254.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S *et al*: **The Vertebrate Genome Annotation (Vega) database**. *Nucleic Acids Res* 2005, **33**(Database Issue):D459-465.

Bai C, Sen P, Hofmann K, Ma L, Goebl M, Harper JW, Elledge SJ: **SKP1 connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box**. *Cell* 1996, **86**(2):263-274.

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M *et al*: **The Universal Protein Resource (UniProt)**. *Nucleic Acids Res* 2005, **33**(Database issue):D154-159.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32 Database issue**:D138-141.

Baumeister W, Walz J, Zuhl F, Seemuller E: **The proteasome: paradigm of a self-compartmentalizing protease**. *Cell* 1998, **92**(3):367-380.

Beatrix B, Sakai H, Wiedmann M: **The alpha and beta subunit of the nascent polypeptide-associated complex have distinct functions**. *J Biol Chem* 2000, **275**(48):37838-37845.

Begley TP, Downs DM, Ealick SE, McLafferty FW, Van Loon AP, Taylor S, Campobasso N, Chiu HJ, Kinsland C, Reddick JJ *et al*: **Thiamin biosynthesis in prokaryotes**. *Arch Microbiol* 1999, **171**(5):293-300.

Benni ML, Neigeborn L: **Identification of a new class of negative regulators affecting sporulation-specific gene expression in yeast**. *Genetics* 1997, **147**(3):1351-1366.

Benton D: **Recent changes in the GenBank On-line Service**. *Nucleic Acids Research* 1990, **18**(6):1517-1520.

Berbee ML, Carmean DA, Winka K: **Ribosomal DNA and resolution of branching order among the ascomycota: how many nucleotides are enough?** *Mol Phylogenet Evol* 2000, **17**(3):337-344.

Bernier-Villamor V, Sampson DA, Matunis MJ, Lima CD: **Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1**. *Cell* 2002, **108**(3):345-356.

Bertolaet BL, Clarke DJ, Wolff M, Watson MH, Henze M, Divita G, Reed SI: **UBA domains of DNA damage-inducible proteins interact with ubiquitin**. *Nat Struct Biol* 2001, **8**(5):417-422.

Bhagwat SS, Manning AM, Hoekstra MF, Lewis A: **Gene-regulating protein kinases as important anti-inflammatory targets**. *Drug Discov Today* 1999, **4**(10):472-479.

Biederer T, Volkwein C, Sommer T: **Role of Cue1p in ubiquitination and degradation at the ER surface**. *Science* 1997, **278**(5344):1806-1809.

Biggins S, Bhalla N, Chang A, Smith DL, Murray AW: **Genes involved in sister chromatid separation and segregation in the budding yeast Saccharomyces cerevisiae**. *Genetics* 2001, **159**(2):453-470.

Bilodeau PS, Winistorfer SC, Kearney WR, Robertson AD, Piper RC: **Vps27-Hse1 and ESCRT-I complexes cooperate to increase efficiency of sorting ubiquitinated proteins at the endosome**. *J Cell Biol* 2003, **163**(2):237-243.

Bimston D, Song J, Winchester D, Takayama S, Reed JC, Morimoto RI: **BAG-1, a negative regulator of Hsp70 chaperone activity, uncouples nucleotide hydrolysis from substrate release**. *EMBO J* 1998, **17**(23):6871-6878.

Birrell GW, Giaever G, Chu AM, Davis RW, Brown JM: **A genome-wide screen in Saccharomyces cerevisiae for genes affecting UV radiation sensitivity**. *Proc Natl Acad Sci U S A* 2001, **98**(22):12608-12613.

Bishop N, Horman A, Woodman P: **Mammalian class E vps proteins recognize ubiquitin and act in the removal of endosomal protein-ubiquitin conjugates**. *J Cell Biol* 2002, **157**(1):91-101.

Blondel M, Mann C: **G2 cyclins are required for the degradation of G1 cyclins in yeast**. *Nature* 1996, **384**(6606):279-282.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome**. *Science* 2003, **299**(5611):1391-1394.

Boguski MS, Lowe TM, Tolstoshev CM: **dbEST--database for "expressed sequence tags"**. *Nat Genet* 1993, **4**(4):332-333.

Boname JM, Stevenson PG: **MHC class I ubiquitination by a viral PHD/LAP finger protein**. *Immunity* 2001, **15**(4):627-636.

Borodovsky A, Kessler BM, Casagrande R, Overkleeft HS, Wilkinson KD, Ploegh HL: **A novel active site-directed probe specific for deubiquitylating enzymes reveals proteasome association of USP14**. *EMBO J* 2001, **20**(18):5187-5196.

Botuyan MV, Koth CM, Mer G, Chakrabartty A, Conaway JW, Conaway RC, Edwards AM, Arrowsmith CH, Chazin WJ: **Binding of elongin A or a von Hippel-Lindau peptide stabilizes the structure of yeast elongin C**. *Proc Natl Acad Sci U S A* 1999, **96**(16):9033-9038.

Brachat S, Dietrich FS, Voegeli S, Zhang Z, Stuart L, Lerch A, Gates K, Gaffney T, Philippsen P: **Reinvestigation of the Saccharomyces cerevisiae genome annotation by comparison to the genome of a related fungus: Ashbya gossypii**. *Genome Biol* 2003, **4**(7):1465-6914.

Braun BC, Glickman M, Kraft R, Dahlmann B, Kloetzel PM, Finley D, Schmidt M: **The base of the proteasome regulatory particle exhibits chaperone-like activity**. *Nat Cell Biol* 1999, **1**(4):221-226.

Broomfield S, Chow BL, Xiao W: **MMS2, encoding a ubiquitin-conjugating-enzyme-like protein, is a member of the yeast error-free postreplication repair pathway**. *Proc Natl Acad Sci U S A* 1998, **95**(10):5678-5683.

Buchberger A, Howard MJ, Proctor M, Bycroft M: **The UBX domain: a widespread ubiquitin-like module**. *J Mol Biol* 2001, **307**(1):17-24.

Bucher P, Karplus K, Moeri N, Hofmann K: **A flexible motif search technique based on generalized profiles**. *Comput Chem* 1996, **20**(1):3-23.

Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**(1):78-94.

Burnett B, Li F, Pittman RN: **The polyglutamine neurodegenerative protein ataxin-3 binds polyubiquitylated proteins and has ubiquitin protease activity**. *Hum Mol Genet* 2003, **12**(23):3195-3205.

Burri L, Hockendorff J, Boehm U, Klamp T, Dohmen RJ, Levy F: **Identification and characterization of a mammalian protein interacting with 20S proteasome precursors**. *Proc Natl Acad Sci U S A* 2000, **97**(19):10348-10353.

Capili AD, Edghill EL, Wu K, Borden KL: **Structure of the C-terminal RING finger from a RING-IBR-RING/TRIAD motif reveals a novel zinc-binding domain distinct from a RING**. *J Mol Biol* 2004, **340**(5):1117-1129.

Chang EC, Schwechheimer C: **ZOMES III: the interface between signalling and proteolysis**. *EMBO Rep* 2004, **5**(11):1041-1045.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Res* 2003, **31**(13):3497-3500.

Chow MK, Mackay JP, Whisstock JC, Scanlon MJ, Bottomley SP: **Structural and functional analysis of the Josephin domain of the polyglutamine protein ataxin-3**. *Biochem Biophys Res Commun* 2004, **322**(2):387-394.

Chung KK, Zhang Y, Lim KL, Tanaka Y, Huang H, Gao J, Ross CA, Dawson VL, Dawson TM: **Parkin ubiquitinates the alpha-synuclein-interacting protein, synphilin-1: implications for Lewy-body formation in Parkinson disease**. *Nat Med* 2001, **7**(10):1144-1150.

Ciccarelli FD, Izaurralde E, Bork P: **The PAM domain, a multi-protein complex-associated module with an all-alpha-helix fold**. *BMC Bioinformatics* 2003, **4**(1):64.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in Saccharomyces genomes by phylogenetic footprinting**. *Science* 2003, **301**(5629):71-76.

Connell P, Ballinger CA, Jiang J, Wu Y, Thompson LJ, Hohfeld J, Patterson C: **The co-chaperone CHIP regulates protein triage decisions mediated by heat-shock proteins**. *Nat Cell Biol* 2001, **3**(1):93-96.

Connelly C, Hieter P: **Budding yeast SKP1 encodes an evolutionarily conserved kinetochore protein required for cell cycle progression**. *Cell* 1996, **86**(2):275-285.

Cope GA, Suh GS, Aravind L, Schwarz SE, Zipursky SL, Koonin EV, Deshaies RJ: **Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1**. *Science* 2002, **298**(5593):608-611.

Coscoy L, Ganem D: **PHD domains and E3 ubiquitin ligases: viruses make the connection**. *Trends Cell Biol* 2003, **13**(1):7-12.

Coscoy L, Sanchez DJ, Ganem D: **A novel class of herpesvirus-encoded membrane-bound E3 ubiquitin ligases regulates endocytosis of proteins involved in immune recognition**. *J Cell Biol* 2001, **155**(7):1265-1273.

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **JPred: a consensus secondary structure prediction server**. *Bioinformatics* 1998, **14**(10):892-893.

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system**. *Genome Res* 2004, **14**(5):942-950.

Cyr DM, Hohfeld J, Patterson C: **Protein quality control: U-box-containing E3 ubiquitin ligases join the fold**. *Trends Biochem Sci* 2002, **27**(7):368-375.

D'Andrea LD, Regan L: **TPR proteins: the versatile helix**. *Trends Biochem Sci* 2003, **28**(12):655-662.

Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact**. *Trends Biochem Sci* 1998, **23**(9):324-328.

Daniel JA, Torok MS, Sun ZW, Schieltz D, Allis CD, Yates JR, 3rd, Grant PA: **Deubiquitination of histone H2B by a yeast acetyltransferase complex regulates transcription**. *J Biol Chem* 2004, **279**(3):1867-1871.

Das R, Zhou Z, Reed R: **Functional association of U2 snRNP with the ATP-independent spliceosomal complex E**. *Mol Cell* 2000, **5**(5):779-787.

Dell'Angelica EC, Puertollano R, Mullins C, Aguilar RC, Vargas JD, Hartnell LM, Bonifacino JS: **GGAs: a family of ADP ribosylation factor-binding proteins related to adaptors and associated with the Golgi complex**. *J Cell Biol* 2000, **149**(1):81-94.

Deshaies RJ: **SCF and Cullin/Ring H2-based ubiquitin ligases**. *Annu Rev Cell Dev Biol* 1999, **15**:435-467.

Deveraux Q, Ustrell V, Pickart C, Rechsteiner M: **A 26 S protease subunit that binds ubiquitin conjugates**. *J Biol Chem* 1994, **269**(10):7059-7061.

Di Fiore PP, Polo S, Hofmann K: **Opinion: When ubiquitin meets ubiquitin receptors: a signalling connection**. *Nat Rev Mol Cell Biol* 2003, **4**(6):491-497.

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S *et al*: **The Ashbya gossypii Genome as a Tool for Mapping the Ancient Saccharomyces cerevisiae Genome**. *Science* 2004, **4**:4.

Dittmar GA, Wilkinson CR, Jedrzejewski PT, Finley D: **Role of a ubiquitin-like modification in polarized morphogenesis**. *Science* 2002, **295**(5564):2442-2446.

Dohmen RJ: **SUMO protein modification**. *Biochim Biophys Acta* 2004, **1695**(1-3):113-131.

Dohmen RJ, Madura K, Bartel B, Varshavsky A: **The N-end rule is mediated by the UBC2(RAD6) ubiquitin-conjugating enzyme**. *Proc Natl Acad Sci U S A* 1991, **88**(16):7351-7355.

Donaldson KM, Yin H, Gekakis N, Supek F, Joazeiro CA: **Ubiquitin signals protein trafficking via interaction with a novel ubiquitin binding domain in the membrane fusion regulator, Vps9p**. *Curr Biol* 2003, **13**(3):258-262.

Dougan DA, Reid BG, Horwich AL, Bukau B: **ClpS, a substrate modulator of the ClpAP machine**. *Mol Cell* 2002, **9**(3):673-683.

Dover J, Schneider J, Tawiah-Boateng MA, Wood A, Dean K, Johnston M, Shilatifard A: **Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6**. *J Biol Chem* 2002, **277**(32):28368-28371.

Durbin R, Eddy S, Krogh A, Mitchison G: **Biological sequence analysis.** 1998.

Durocher D, Henckel J, Fersht AR, Jackson SP: **The FHA domain is a modular phosphopeptide recognition motif**. *Mol Cell* 1999, **4**(3):387-394.

Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.

Elsasser S, Chandler-Militello D, Muller B, Hanna J, Finley D: **Rad23 and Rpn10 serve as alternative ubiquitin receptors for the proteasome**. *J Biol Chem* 2004, **279**(26):26817-26822.

Elsasser S, Gali RR, Schwickart M, Larsen CN, Leggett DS, Muller B, Feng MT, Tubing F, Dittmar GA, Finley D: **Proteasome subunit Rpn1 binds ubiquitin-like protein domains**. *Nat Cell Biol* 2002, **4**(9):725-730.

Evans PC, Ovaa H, Hamon M, Kilshaw PJ, Hamm S, Bauer S, Ploegh HL, Smith TS: **Zinc-finger protein A20, a regulator of inflammation and cell survival, has de-ubiquitinating activity**. *Biochem J* 2004, **378**(Pt 3):727-734.

Evans PC, Smith TS, Lai MJ, Williams MG, Burke DF, Heyninck K, Kreike MM, Beyaert R, Blundell TL, Kilshaw PJ: **A novel type of deubiquitinating enzyme**. *J Biol Chem* 2003, **278**(25):23180-23186.

Evans RM, Hollenberg SM: **Zinc fingers: gilt by association**. *Cell* 1988, **52**(1):1-3.

Farras R, Ferrando A, Jasik J, Kleinow T, Okresz L, Tiburcio A, Salchert K, del Pozo C, Schell J, Koncz C: **SKP1-SnRK protein kinase interactions mediate proteasomal binding of a plant SCF ubiquitin ligase**. *EMBO J* 2001, **20**(11):2742-2756.

Ferrell K, Wilkinson CR, Dubiel W, Gordon C: **Regulatory subunit interactions of the 26S proteasome, a complex problem**. *Trends Biochem Sci* 2000, **25**(2):83-88.

Finley D, Bartel B, Varshavsky A: **The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome biogenesis**. *Nature* 1989, **338**(6214):394-401.

Fitch WM: **Homology a personal view on some of the problems**. *Trends Genet* 2000, **16**(5):227-231.

Foury F: **Human genetic diseases: a cross-talk between man and yeast**. *Gene* 1997, **195**(1):1-10.

Fraschini R, Bilotta D, Lucchini G, Piatti S: **Functional characterization of Dma1 and Dma2, the budding yeast homologues of Schizosaccharomyces pombe Dma1 and human Chfr**. *Mol Biol Cell* 2004, **15**(8):3796-3810.

Freilich S, Oron E, Kapp Y, Nevo-Caspi Y, Orgad S, Segal D, Chamovitz DA: **The COP9 signalosome is essential for development of Drosophila melanogaster**. *Curr Biol* 1999, **9**(20):1187-1190.

Fruh K, Bartee E, Gouveia K, Mansouri M: **Immune evasion by a novel family of viral PHD/LAP-finger proteins of gamma-2 herpesviruses and poxviruses**. *Virus Res* 2002, **88**(1-2):55-69.

Furukawa K, Mizushima N, Noda T, Ohsumi Y: **A protein conjugation system in yeast with homology to biosynthetic enzyme reaction of prokaryotes**. *J Biol Chem* 2000, **275**(11):7462-7465.

Galan JM, Wiederkehr A, Seol JH, Haguenauer-Tsapis R, Deshaies RJ, Riezman H, Peter M: **Skp1p and the F-box protein Rcy1p form a non-SCF complex involved in recycling of the SNARE Snc1p in yeast**. *Mol Cell Biol* 2001, **21**(9):3105-3117.

Gan-Erdene T, Nagamalleswari K, Yin L, Wu K, Pan ZQ, Wilkinson KD: **Identification and characterization of DEN1, a deneddylase of the ULP family**. *J Biol Chem* 2003, **278**(31):28892-28900.

Gandre S, Kahana C: **Degradation of ornithine decarboxylase in Saccharomyces cerevisiae is ubiquitin independent**. *Biochem Biophys Res Commun* 2002, **293**(1):139-144.

Garrus JE, von Schwedler UK, Pornillos OW, Morham SG, Zavitz KH, Wang HE, Wettstein DA, Stray KM, Cote M, Rich RL *et al*: **Tsg101 and the vacuolar protein sorting pathway are essential for HIV-1 budding**. *Cell* 2001, **107**(1):55-65.

Gavin AC, Bosche M, Krause R, Merino A, Drewes G, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G *et al*: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**(6868):141-147.

Gille C, Goede A, Schloetelburg C, Preissner R, Kloetzel PM, Gobel UB, Frommel C: **A comprehensive view on proteasomal sequences: implications for the evolution of the proteasome**. *J Mol Biol* 2003, **326**(5):1437-1448.

Glickman MH, Rubin DM, Coux O, Wefes I, Pfeifer G, Cjeka Z, Baumeister W, Fried VA, Finley D: **A subcomplex of the proteasome regulatory particle required for ubiquitin-conjugate degradation and related to the COP9-signalosome and eIF3**. *Cell* 1998, **94**(5):615-623.

Goehring AS, Rivers DM, Sprague GF, Jr.: **Urmylation: a ubiquitin-like pathway that functions during invasive growth and budding in yeast**. *Mol Biol Cell* 2003, **14**(11):4329-4341.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al*: **Life with 6000 genes**. *Science* 1996, **274**(5287):546, 563-547.

Goh PY, Surana U: **Cdc4, a protein required for the onset of S phase, serves an essential function during G(2)/M transition in Saccharomyces cerevisiae**. *Mol Cell Biol* 1999, **19**(8):5512-5522.

Gong L, Li B, Millas S, Yeh ET: **Molecular cloning and characterization of human AOS1 and UBA2, components of the sentrin-activating enzyme complex**. *FEBS Lett* 1999, **448**(1):185-189.

Gostissa M, Hengstermann A, Fogal V, Sandy P, Schwarz SE, Scheffner M, Del Sal G: **Activation of p53 by conjugation to the ubiquitin-like protein SUMO-1**. *EMBO J* 1999, **18**(22):6462-6471.

Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 2001, **313**(4):903-919.

Grant RP, Neuhaus D, Stewart M: **Structural basis for the interaction between the Tap/NXF1 UBA domain and FG nucleoporins at 1A resolution**. *J Mol Biol* 2003, **326**(3):849-858.

Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins**. *Proc Natl Acad Sci U S A* 1987, **84**(13):4355-4358.

Groisman R, Polanowska J, Kuraoka I, Sawada J, Saijo M, Drapkin R, Kisselev AF, Tanaka K, Nakatani Y: **The ubiquitin ligase activity in the DDB2 and CSA complexes is differentially regulated by the COP9 signalosome in response to DNA damage**. *Cell* 2003, **113**(3):357-367.

Groll M, Bajorek M, Kohler A, Moroder L, Rubin DM, Huber R, Glickman MH, Finley D: **A gated channel into the proteasome core particle**. *Nat Struct Biol* 2000, **7**(11):1062-1067.

Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling**. *Electrophoresis* 1997, **18**(15):2714-2723.

Gusella JF, MacDonald ME: **Molecular genetics: unmasking polyglutamine triggers in neurodegenerative disease**. *Nat Rev Neurosci* 2000, **1**(2):109-115.

Haas AL, Siepmann TJ: **Pathways of ubiquitin conjugation**. *FASEB J* 1997, **11**(14):1257-1268.

Hartmann-Petersen R, Semple CA, Ponting CP, Hendil KB, Gordon C: **UBA domain containing proteins in fission yeast**. *Int J Biochem Cell Biol* 2003, **35**(5):629-636.

Hartmann-Petersen R, Wallace M, Hofmann K, Koch G, Johnsen AH, Hendil KB, Gordon C: **The Ubx2 and Ubx3 cofactors direct Cdc48 activity to proteolytic and nonproteolytic ubiquitin-dependent processes**. *Curr Biol* 2004, **14**(9):824-828.

Hatakeyama S, Yada M, Matsumoto M, Ishida N, Nakayama KI: **U box proteins as a new family of ubiquitin-protein ligases**. *J Biol Chem* 2001, **276**(35):33111-33120.

Hauser HP, Bardroff M, Pyrowolakis G, Jentsch S: **A giant ubiquitin-conjugating enzyme related to IAP apoptosis inhibitors**. *J Cell Biol* 1998, **141**(6):1415-1422.

Hendil KB, Khan S, Tanaka K: **Simultaneous binding of PA28 and PA700 activators to 20 S proteasomes**. *Biochem J* 1998, **332**(Pt 3):749-754.

Henikoff S, Henikoff JG: **Automated assembly of protein blocks for database searching**. *Nucleic Acids Res* 1991, **19**(23):6565-6572.

Hershko A, Ciechanover A: **The ubiquitin system**. 1998:425-479.

Hershko A, Heller H, Elias S, Ciechanover A: **Components of ubiquitin-protein ligase system. Resolution, affinity purification, and role in protein breakdown**. *J Biol Chem* 1983, **258**(13):8206-8214.

Hetfeld BKJ, Helfrich A, Kapelari B, Scheel H, Hofmann K, Guterman A, Glickman M, Dubiel W, Schade R, Kloetzel P: **The COP9 signalosome associated deubiquitinating enzyme USP15 exhibits a Zinc-finger that is essential for rescuing the ubiquitin ligase Rbx1 from degradation.** *Curr Biol.* 2005 Jul 12;15(13):1217-21.

Hicke L: **Protein regulation by monoubiquitin**. *Nat Rev Mol Cell Biol* 2001, **2**(3):195-201.

Hipp MS, Raasi S, Groettrup M, Schmidtke G: **NEDD8 ultimate buster-1L interacts with the ubiquitin-like protein FAT10 and accelerates its degradation**. *J Biol Chem* 2004, **279**(16):16503-16510.

Hiyama H, Yokoi M, Masutani C, Sugasawa K, Maekawa T, Tanaka K, Hoeijmakers JH, Hanaoka F: **Interaction of hHR23 with S5a. The ubiquitin-like domain of hHR23 mediates interaction with S5a subunit of 26 S proteasome**. *J Biol Chem* 1999, **274**(39):28019-28025.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al*: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry**. *Nature* 2002, **415**(6868):180-183.

Hoege C, Pfander B, Moldovan GL, Pyrowolakis G, Jentsch S: **RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO**. *Nature* 2002, **419**(6903):135-141.

Hofmann K: **Sensitive protein comparisons with profiles and hidden Markov models**. *Brief Bioinform* 2000, **1**(2):167-178.

Hofmann K, Bucher P: **The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors**. *Trends Biochem Sci* 1995, **20**(9):347-349.

Hofmann K, Bucher P: **The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway**. *Trends Biochem Sci* 1996, **21**(5):172-173.

Hofmann K, Bucher P: **The PCI domain: a common theme in three multiprotein complexes**. *Trends Biochem Sci* 1998, **23**(6):204-205.

Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999**. *Nucleic Acids Res* 1999, **27**(1):215-219.

Hofmann K, Falquet L: **A ubiquitin-interacting motif conserved in components of the proteasomal and lysosomal protein degradation systems**. *Trends Biochem Sci* 2001, **26**(6):347-350.

Hofmann RM, Pickart CM: **Noncanonical MMS2-encoded ubiquitin-conjugating enzyme functions in assembly of novel polyubiquitin chains for DNA repair**. *Cell* 1999, **96**(5):645-653.

Holm L, Sander C: **Dali/FSSP classification of three-dimensional protein folds**. *Nucleic Acids Res* 1997, **25**(1):231-234.

Hook SS, Orian A, Cowley SM, Eisenman RN: **Histone deacetylase 6 binds polyubiquitin through its zinc finger (PAZ domain) and copurifies with deubiquitinating enzymes**. *Proc Natl Acad Sci U S A* 2002, **99**(21):13425-13430.

Hu M, Li P, Li M, Li W, Yao T, Wu JW, Gu W, Cohen RE, Shi Y: **Crystal structure of a UBP-family deubiquitinating enzyme in isolation and in complex with ubiquitin aldehyde**. *Cell* 2002, **111**(7):1041-1054.

Huang DT, Miller DW: **A time-efficient, linear-space local similarity algorithm.** *Adv Appl Math* 1991, **12**:337-357.

Huang DT, Paydar A, Zhuang M, Waddell MB, Holton JM, Schulman BA: **Structural basis for recruitment of Ubc12 by an E2 binding domain in NEDD8's E1**. *Mol Cell* 2005, **17**(3):341-350.

Huang DT, Walden H, Duda D, Schulman BA: **Ubiquitin-like protein activation**. *Oncogene* 2004, **23**(11):1958-1971.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Stupka E, Ureta-Vidal A, Vastrik I *et al*: **The Ensembl genome database project**. *Nucleic Acids Res* 2002, **30**(1):38-41.

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast**. *Nature* 2003, **425**(6959):686-691.

Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, Bairoch A: **Recent improvements to the PROSITE database**. *Nucleic Acids Res* 2004, **32 Database issue**:D134-137.

Ideguchi H, Ueda A, Tanaka M, Yang J, Tsuji T, Ohno S, Hagiwara E, Aoki A, Ishigatsubo Y: **Structural and functional characterization of the USP11 deubiquitinating enzyme, which interacts with the RanGTP-associated protein RanBPM**. *Biochem J* 2002, **367**(Pt 1):87-95.

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292**(5518):929-934.

Imai J, Maruya M, Yashiroda H, Yahara I, Tanaka K: **The molecular chaperone Hsp90 plays a role in the assembly and maintenance of the 26S proteasome**. *EMBO J* 2003, **22**(14):3557-3567.

Isono E, Saeki Y, Yokosawa H, Toh-e A: **Rpn7 Is required for the structural integrity of the 26 S proteasome of Saccharomyces cerevisiae**. *J Biol Chem* 2004, **279**(26):27168-27176.

Issel-Tarver L, Christie KR, Dolinski K, Andrada R, Balakrishnan R, Ball CA, Binkley G, Dong S, Dwight SS, Fisk DG *et al*: **Saccharomyces Genome Database**. 2002:329-346.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proc Natl Acad Sci U S A* 2001, **98**(8):4569-4574.

Jensen DE, Proctor M, Marquis ST, Gardner HP, Ha SI, Chodosh LA, Ishov AM, Tommerup N, Vissing H, Sekido Y *et al*: **BAP1: a novel ubiquitin hydrolase which binds to the BRCA1 RING finger and enhances BRCA1-mediated cell growth suppression**. *Oncogene* 1998, **16**(9):1097-1112.

Jentsch S, Pyrowolakis G: **Ubiquitin and its kin: how close are the family ties?** *Trends Cell Biol* 2000, **10**(8):335-342.

Jiang J, Ballinger CA, Wu Y, Dai Q, Cyr DM, Hohfeld J, Patterson C: **CHIP is a U-box-dependent E3 ubiquitin ligase: identification of Hsc70 as a target for ubiquitylation**. *J Biol Chem* 2001, **276**(46):42938-42944.

Johnson ES, Blobel G: **Ubc9p is the conjugating enzyme for the ubiquitin-like protein Smt3p**. *J Biol Chem* 1997, **272**(43):26799-26802.

Johnson ES, Gupta AA: **An E3-like factor that promotes SUMO conjugation to the yeast septins**. *Cell* 2001, **106**(6):735-744.

Johnson ES, Schwienhorst I, Dohmen RJ, Blobel G: **The ubiquitin-like protein Smt3p is activated for conjugation to other proteins by an Aos1p/Uba2p heterodimer**. *EMBO J* 1997, **16**(18):5509-5519.

Jones D, Crowe E, Stevens TA, Candido EP: **Functional and phylogenetic analysis of the ubiquitylation system in Caenorhabditis elegans: ubiquitin-conjugating enzymes, ubiquitin-activating enzymes, and ubiquitin-like proteins**. *Genome Biol 2002* 2001, **3**(1):1465-6914.

Jones S, Thornton JM: **Protein-protein interactions: a review of protein dimer structures**. 1995:31-65.

Jones S, Thornton JM: **Principles of protein-protein interactions**. *Proc Natl Acad Sci U S A* 1996, **93**(1):13-20.

Kagey MH, Melhuish TA, Wotton D: **The polycomb protein Pc2 is a SUMO E3**. *Cell* 2003, **113**(1):127-137.

Kajava AV: **What curves alpha-solenoids? Evidence for an alpha-helical toroid structure of Rpn1 and Rpn2 proteins of the 26 S proteasome**. *J Biol Chem* 2002, **277**(51):49791-49798.

Kang RS, Daniels CM, Francis SA, Shih SC, Salerno WJ, Hicke L, Radhakrishnan I: **Solution structure of a CUE-ubiquitin complex reveals a conserved mode of ubiquitin binding**. *Cell* 2003, **113**(5):621-630.

Kapelari B, Bech-Otschir D, Hegerl R, Schade R, Dumdey R, Dubiel W: **Electron microscopy and subunit-subunit interaction studies reveal a first architecture of COP9 signalosome**. *J Mol Biol* 2000, **300**(5):1169-1178.

Kaplun L, Ivantsiv Y, Bakhrat A, Raveh D: **DNA damage response-mediated degradation of Ho endonuclease via the ubiquitin system involves its nuclear export**. *J Biol Chem* 2003, **278**(49):48727-48734.

Kas K, Stickens D, Merregaert J: **Characterization of a processed pseudogene of human FAU1 on chromosome 18**. *Gene* 1995, **160**(2):273-276.

Katoh Y, Shiba Y, Mitsuhashi H, Yanagida Y, Takatsu H, Nakayama K: **Tollip and Tom1 form a complex and recruit ubiquitin-conjugated proteins onto early endosomes**. *J Biol Chem* 2004, **279**(23):24435-24443.

Katzmann DJ, Babst M, Emr SD: **Ubiquitin-dependent sorting into the multivesicular body pathway requires the function of a conserved endosomal protein sorting complex, ESCRT-I**. *Cell* 2001, **106**(2):145-155.

Katzmann DJ, Odorizzi G, Emr SD: **Receptor downregulation and multivesicular-body sorting**. *Nat Rev Mol Cell Biol* 2002, **3**(12):893-905.

Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, Kawakami H, Nakamura S, Nishimura M, Akiguchi I: **CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1**. *Nat Genet* 1994, **8**(3):221-228.

Kawashima T, Berthet-Colominas C, Wulff M, Cusack S, Leberman R: **The structure of the Escherichia coli EF-Tu.EF-Ts complex at 2.5 A resolution**. *Nature* 1996, **379**(6565):511-518.

Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae**. *Nature* 2004, **428**(6983):617-624.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements**. *Nature* 2003, **423**(6937):241-254.

Kim KI, Giannakopoulos NV, Virgin HW, Zhang DE: **Interferon-inducible ubiquitin E2, Ubc8, is a conjugating enzyme for protein ISGylation**. *Mol Cell Biol* 2004, **24**(21):9592-9600.

Kim KI, Zhang DE: **ISG15, not just another ubiquitin-like protein**. *Biochem Biophys Res Commun* 2003, **307**(3):431-434.

Kitada T, Asakawa S, Hattori N, Matsumine H, Yamamura Y, Minoshima S, Yokochi M, Mizuno Y, Shimizu N: **Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism**. *Nature* 1998, **392**(6676):605-608.

Klapisz E, Sorokina I, Lemeer S, Pijnenburg M, Verkleij AJ, van Bergen en Henegouwen PM: **A ubiquitin-interacting motif (UIM) is essential for Eps15 and Eps15R ubiquitination**. *J Biol Chem* 2002, **277**(34):30746-30753.

Kloetzel PM: **Generation of major histocompatibility complex class I antigens: functional interplay between proteasomes and TPPII**. *Nat Immunol* 2004, **5**(7):661-669.

Kloetzel PM: **The proteasome and MHC class I antigen processing**. *Biochim Biophys Acta* 2004, **1695**(1-3):225-233.

Koegl M, Hoppe T, Schlenker S, Ulrich HD, Mayer TU, Jentsch S: **A novel ubiquitination factor, E4, is involved in multiubiquitin chain assembly**. *Cell* 1999, **96**(5):635-644.

Komatsu M, Chiba T, Tatsumi K, Iemura S, Tanida I, Okazaki N, Ueno T, Kominami E, Natsume T, Tanaka K: **A novel protein-conjugating system for Ufm1, a ubiquitin-fold modifier**. *EMBO J* 2004, **23**(9):1977-1986.

Komatsu M, Tanida I, Ueno T, Ohsumi M, Ohsumi Y, Kominami E: **The C-terminal region of an Apg7p/Cvt2p is required for homodimerization and is essential for its E1 activity and E1-E2 complex formation**. *J Biol Chem* 2001, **276**(13):9846-9854.

Kostriken R, Strathern JN, Klar AJ, Hicks JB, Heffron F: **A site-specific endonuclease essential for mating-type switching in Saccharomyces cerevisiae**. *Cell* 1983, **35**(1):167-174.

Kumar S, Talis AL, Howley PM: **Identification of HHR23A as a substrate for E6-associated protein-mediated ubiquitination**. *J Biol Chem* 1999, **274**(26):18785-18792.

Kwon J, Wang YL, Setsuie R, Sekiguchi S, Sato Y, Sakurai M, Noda M, Aoki S, Yoshikawa Y, Wada K: **Two closely related ubiquitin C-terminal hydrolase isozymes function as reciprocal modulators of germ cell apoptosis in cryptorchid testis**. *Am J Pathol* 2004, **165**(4):1367-1374.

Kwon YT, Reiss Y, Fried VA, Hershko A, Yoon JK, Gonda DK, Sangan P, Copeland NG, Jenkins NA, Varshavsky A: **The mouse and human genes encoding the recognition component of the N-end rule pathway**. *Proc Natl Acad Sci U S A* 1998, **95**(14):7898-7903.

Lake MW, Wuebbens MM, Rajagopalan KV, Schindelin H: **Mechanism of ubiquitin activation revealed by the structure of a bacterial MoeB-MoaD complex**. *Nature* 2001, **414**(6861):325-329.

Lam YA, DeMartino GN, Pickart CM, Cohen RE: **Specificity of the ubiquitin isopeptidase in the PA700 regulatory complex of 26 S proteasomes**. *J Biol Chem* 1997, **272**(45):28438-28446.

Lam YA, Lawson TG, Velayutham M, Zweier JL, Pickart CM: **A proteasomal ATPase subunit recognizes the polyubiquitin degradation signal**. *Nature* 2002, **416**(6882):763-767.

Lam YA, Xu W, DeMartino GN, Cohen RE: **Editing of ubiquitin conjugates by an isopeptidase in the 26S proteasome**. *Nature* 1997, **385**(6618):737-740.

Lammer D, Mathias N, Laplaza JM, Jiang W, Liu Y, Callis J, Goebl M, Estelle M: **Modification of yeast Cdc53p by the ubiquitin-related protein rub1p affects function of the SCFCdc4 complex**. *Genes Dev* 1998, **12**(7):914-926.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

Leggett DS, Hanna J, Borodovsky A, Crosas B, Schmidt M, Baker RT, Walz T, Ploegh H, Finley D: **Multiple associated proteins regulate proteasome structure and function**. *Mol Cell* 2002, **10**(3):495-507.

Leroy E, Boyer R, Auburger G, Leube B, Ulm G, Mezey E, Harta G, Brownstein MJ, Jonnalagada S, Chernova T *et al*: **The ubiquitin pathway in Parkinson's disease**. *Nature* 1998, **395**(6701):451-452.

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration**. *Nucleic Acids Res* 2004, **32 Database issue**:D142-144.

Li L, Li X, Francke U, Cohen SN: **The TSG101 tumor susceptibility gene is located in chromosome 11 band p15 and is mutated in human breast cancer**. *Cell* 1997, **88**(1):143-154.

Li M, Brooks CL, Wu-Baer F, Chen D, Baer R, Gu W: **Mono- versus polyubiquitination: differential control of p53 fate by Mdm2**. *Science* 2003, **302**(5652):1972-1975.

Li SJ, Hochstrasser M: **The Ulp1 SUMO isopeptidase: distinct domains required for viability, nuclear envelope localization, and substrate specificity**. *J Cell Biol* 2003, **160**(7):1069-1081.

Li WH: **Molecular Evolution**. 1997.

Li X, Coffino P: **Distinct domains of antizyme required for binding and proteolysis of ornithine decarboxylase**. *Mol Cell Biol* 1994, **14**(1):87-92.

Lier S, Paululat A: **The proteasome regulatory particle subunit Rpn6 is required for Drosophila development and interacts physically with signalosome subunit Alien/CSN2**. *Gene* 2002, **298**(2):109-119.

Linghu B, Callis J, Goebl MG: **Rub1p processing by Yuh1p is required for wild-type levels of Rub1p conjugation to Cdc53p**. *Eukaryot Cell* 2002, **1**(3):491-494.

Lipman DJ, Altschul SF, Kececioglu JD: **A tool for multiple sequence alignment**. *Proc Natl Acad Sci U S A* 1989, **86**(12):4412-4415.

Lois LM, Lima CD: **Structures of the SUMO E1 provide mechanistic insights into SUMO activation and E2 recruitment to E1**. 2005.

Lu Z, Xu S, Joazeiro C, Cobb MH, Hunter T: **The PHD domain of MEKK1 acts as an E3 ubiquitin ligase and mediates ubiquitination and degradation of ERK1/2**. *Mol Cell* 2002, **9**(5):945-956.

Luders J, Demand J, Hohfeld J: **The ubiquitin-related BAG-1 provides a link between the molecular chaperones Hsc70/Hsp70 and the proteasome**. *J Biol Chem* 2000, **275**(7):4613-4617.

Luders J, Pyrowolakis G, Jentsch S: **The ubiquitin-like protein HUB1 forms SDS-resistant complexes with cellular proteins in the absence of ATP**. *EMBO Rep* 2003, **4**(12):1169-1174.

Lupas AN, Koretke KK: **Bioinformatic analysis of ClpS, a protein module involved in prokaryotic and eukaryotic protein degradation**. *J Struct Biol* 2003, **141**(1):77-83.

Ma J, Katz E, Belote JM: **Expression of proteasome subunit isoforms during spermatogenesis in Drosophila melanogaster**. *Insect Mol Biol* 2002, **11**(6):627-639.

Makarova KS, Aravind L, Koonin EV: **A novel superfamily of predicted cysteine proteases from eukaryotes, viruses and Chlamydia pneumoniae**. *Trends Biochem Sci* 2000, **25**(2):50-52.

Mannhaupt G, Schnall R, Karpov V, Vetter I, Feldmann H: **Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast**. *FEBS Lett* 1999, **450**(1-2):27-34.

Mansouri M, Bartee E, Gouveia K, Hovey Nerenberg BT, Barrett J, Thomas L, Thomas G, McFadden G, Fruh K: **The PHD/LAP-domain protein M153R of myxomavirus is a ubiquitin ligase that induces the rapid internalization and lysosomal destruction of CD4**. *J Virol* 2003, **77**(2):1427-1440.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function**. *Nature* 1999, **402**(6757):83-86.

Margolis RL, Ross CA: **Expansion explosion: new clues to the pathogenesis of repeat expansion neurodegenerative diseases**. *Trends Mol Med* 2001, **7**(11):479-482.

Marin I, Lucas JI, Gradilla AC, Ferrus A: **Parkin and relatives: the RBR family of ubiquitin ligases**. *Physiol Genomics* 2004, **17**(3):253-263.

Matheny SA, Chen C, Kortum RL, Razidlo GL, Lewis RE, White MA: **Ras regulates assembly of mitogenic signalling complexes through the effector protein IMP**. *Nature* 2004, **427**(6971):256-260.

Mathias N, Steussy CN, Goebl MG: **An essential domain within Cdc34p is required for binding to a complex containing Cdc4p and Cdc53p in Saccharomyces cerevisiae**. *J Biol Chem* 1998, **273**(7):4040-4045.

Matsufuji S, Matsufuji T, Miyazaki Y, Murakami Y, Atkins JF, Gesteland RF, Hayashi S: **Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme**. *Cell* 1995, **80**(1):51-60.

Matunis MJ, Coutavas E, Blobel G: **A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex**. *J Cell Biol* 1996, **135**(6 Pt 1):1457-1470.

Mayeur GL, Fraser CS, Peiretti F, Block KL, Hershey JW: **Characterization of eIF3k: a newly discovered subunit of mammalian translation initiation factor eIF3**. *Eur J Biochem* 2003, **270**(20):4133-4139.

Maytal-Kivity V, Pick E, Piran R, Hofmann K, Glickman MH: **The COP9 signalosome-like complex in S. cerevisiae and links to other PCI complexes**. *Int J Biochem Cell Biol* 2003, **35**(5):706-715.

Maytal-Kivity V, Reis N, Hofmann K, Glickman MH: **MPN+, a putative catalytic motif found in a subset of MPN domain proteins from eukaryotes and prokaryotes, is critical for Rpn11 function**. *BMC Biochem* 2002, **3**(1):28.

McCullough J, Clague MJ, Urbe S: **AMSH is an endosome-associated ubiquitin isopeptidase**. *J Cell Biol* 2004, **166**(4):487-492.

Melchior F, Schergaut M, Pichler A: **SUMO: ligases, isopeptidases and nuclear pores**. *Trends Biochem Sci* 2003, **28**(11):612-618.

Meyer HH, Wang Y, Warren G: **Direct binding of ubiquitin conjugates by the mammalian p97 adaptor complexes, p47 and Ufd1-Npl4**. *Embo J* 2002, **21**(21):5645-5652.

Michiels L, Van der Rauwelaert E, Van Hasselt F, Kas K, Merregaert J: **fau cDNA encodes a ubiquitin-like-S30 fusion protein and is expressed as an antisense sequence in the Finkel-Biskis-Reilly murine sarcoma virus**. *Oncogene* 1993, **8**(9):2537-2546.

Minty A, Dumont X, Kaghad M, Caput D: **Covalent modification of p73alpha by SUMO-1. Two-hybrid screening with p73 identifies novel SUMO-1-interacting proteins and a SUMO-1 interaction motif**. *J Biol Chem* 2000, **275**(46):36316-36323.

Mizushima N, Yoshimori T, Ohsumi Y: **Role of the Apg12 conjugation system in mammalian autophagy**. *Int J Biochem Cell Biol* 2003, **35**(5):553-561.

Moazed D, Johnson D: **A deubiquitinating enzyme interacts with SIR4 and regulates silencing in S. cerevisiae**. *Cell* 1996, **86**(4):667-677.

Morris-Desbois C, Rety S, Ferro M, Garin J, Jalinot P: **The human protein HSPC021 interacts with Int-6 and is associated with eukaryotic translation initiation factor 3**. *J Biol Chem* 2001, **276**(49):45988-45995.

Moynihan TP, Ardley HC, Nuber U, Rose SA, Jones PF, Markham AF, Scheffner M, Robinson PA: **The ubiquitin-conjugating enzymes UbcH7 and UbcH8 interact with RING finger/IBR motif-containing domains of HHARI and H7-AP1**. *J Biol Chem* 1999, **274**(43):30963-30968.

Mueller TD, Feigon J: **Solution structures of UBA domains reveal a conserved hydrophobic surface for protein-protein interactions**. *J Mol Biol* 2002, **319**(5):1243-1255.

Muller S, Hoege C, Pyrowolakis G, Jentsch S: **SUMO, ubiquitin's mysterious cousin**. *Nat Rev Mol Cell Biol* 2001, **2**(3):202-210.

Murakami Y, Matsufuji S, Kameji T, Hayashi S, Igarashi K, Tamura T, Tanaka K, Ichihara A: **Ornithine decarboxylase is degraded by the 26S proteasome without ubiquitination**. *Nature* 1992, **360**(6404):597-599.

Nakamura M, Tanigawa Y: **Characterization of ubiquitin-like polypeptide acceptor protein, a novel pro-apoptotic member of the Bcl2 family**. *Eur J Biochem* 2003, **270**(20):4052-4058.

Namy O, Rousset JP, Napthine S, Brierley I: **Reprogrammed genetic decoding in cellular gene expression**. *Mol Cell* 2004, **13**(2):157-168.

Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment**. *J Mol Biol* 2000, **302**(1):205-217.

O'Neill BM, Hanway D, Winzeler EA, Romesberg FE: **Coordinated functions of WSS1, PSY2 and TOF1 in the DNA damage response**. *Nucleic Acids Res* 2004, **32**(22):6519-6530. Print 2004.

Ohh M, Kim WY, Moslehi JJ, Chen Y, Chau V, Read MA, Kaelin WG, Jr.: **An intact NEDD8 pathway is required for Cullin-dependent ubiquitylation in mammalian cells**. *EMBO Rep* 2002, **3**(2):177-182.

Ohi MD, Vander Kooi CW, Rosenberg JA, Chazin WJ, Gould KL: **Structural insights into the U-box, a domain associated with multi-ubiquitination**. *Nat Struct Biol* 2003, **10**(4):250-255.

Ohsumi Y: **Molecular dissection of autophagy: two ubiquitin-like systems**. *Nat Rev Mol Cell Biol* 2001, **2**(3):211-216.

Ooi SL, Shoemaker DD, Boeke JD: **A DNA microarray-based genetic screen for nonhomologous end-joining mutants in Saccharomyces cerevisiae**. *Science* 2001, **294**(5551):2552-2556.

Ozkaynak E, Finley D, Solomon MJ, Varshavsky A: **The yeast ubiquitin genes: a family of natural gene fusions**. *EMBO J* 1987, **6**(5):1429-1439.

Ozkaynak E, Finley D, Varshavsky A: **The yeast ubiquitin gene: head-to-tail repeats encoding a polyubiquitin precursor protein**. *Nature* 1984, **312**(5995):663-666.

Palanimurugan R, Scheel H, Hofmann K, Dohmen RJ: **Polyamines regulate their synthesis by inducing expression and blocking degradation of ODC antizyme**. *EMBO J* 2004, **23**(24):4857-4867.

Palenchar PM, Buck CJ, Cheng H, Larson TJ, Mueller EG: **Evidence that ThiI, an enzyme shared between thiamin and 4-thiouridine biosynthesis, may be a sulfurtransferase that proceeds through a persulfide intermediate**. *J Biol Chem* 2000, **275**(12):8283-8286.

Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison**. *Proc Natl Acad Sci U S A* 1988, **85**(8):2444-2448.

Perez MK, Paulson HL, Pittman RN: **Ataxin-3 with an altered conformation that exposes the polyglutamine domain is associated with the nuclear matrix**. *Hum Mol Genet* 2003, **8**(13):2377-2385.

Peters JM: **The anaphase-promoting complex: proteolysis in mitosis and beyond**. *Mol Cell* 2002, **9**(5):931-943.

Pichler A, Gast A, Seeler JS, Dejean A, Melchior F: **The nucleoporin RanBP2 has SUMO1 E3 ligase activity**. *Cell* 2002, **108**(1):109-120.
Pickart CM: **Ubiquitin in chains**. *Trends Biochem Sci* 2000, **25**(11):544-548.
Pickart CM: **Mechanisms underlying ubiquitination**. 2001:503-533.
Pickart CM, Eddins MJ: **Ubiquitin: structures, functions, mechanisms**. *Biochim Biophys Acta* 2004, **1695**(1-3):55-72.
Pickart CM, Rose IA: **Functional heterogeneity of ubiquitin carrier proteins**. *J Biol Chem* 1985, **260**(3):1573-1581.
Pintard L, Willis JH, Willems A, Johnson JL, Srayko M, Kurz T, Glaser S, Mains PE, Tyers M, Bowerman B *et al*: **The BTB protein MEL-26 is a substrate-specific adaptor of the CUL-3 ubiquitin-ligase**. *Nature* 2003, **425**(6955):311-316.
Pitterle DM, Rajagopalan KV: **The biosynthesis of molybdopterin in Escherichia coli. Purification and characterization of the converting factor**. *J Biol Chem* 1993, **268**(18):13499-13505.
Polo S, Sigismund S, Faretta M, Guidi M, Capua MR, Bossi G, Chen H, De Camilli P, Di Fiore PP: **A single motif responsible for ubiquitin recognition and monoubiquitination in endocytic proteins**. *Nature* 2002, **416**(6879):451-455.
Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, Pike B, Root H, Rubenstein J, Boyer R *et al*: **Mutation in the alpha-synuclein gene identified in families with Parkinson's disease**. *Science* 1997, **276**(5321):2045-2047.
Ponting CP: **Proteins of the endoplasmic-reticulum-associated degradation pathway: domain detection and function prediction**. *Biochem J* 2000, **351 Pt 2**:527-535.
Ponting CP, Cai YD, Bork P: **The breast cancer gene product TSG101: a regulator of ubiquitination?** *J Mol Med* 1997, **75**(7):467-469.
Pornillos O, Alam SL, Davis DR, Sundquist WI: **Structure of the Tsg101 UEV domain in complex with the PTAP motif of the HIV-1 p6 protein**. *Nat Struct Biol* 2002, **9**(11):812-817.
Pornillos O, Alam SL, Rich RL, Myszka DG, Davis DR, Sundquist WI: **Structure and functional interactions of the Tsg101 UEV domain**. *Embo J* 2002, **21**(10):2397-2406.
Prakash S, Prakash L: **Nucleotide excision repair in yeast**. *Mutat Res* 2000, **451**(1-2):13-24.
Pringa E, Martinez-Noel G, Muller U, Harbers K: **Interaction of the ring finger-related U-box motif of a nuclear dot protein with ubiquitin-conjugating enzymes**. *J Biol Chem* 2001, **276**(22):19617-19623.
Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A *et al*: **ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins**. *Nucleic Acids Res* 2003, **31**(13):3625-3630.
Raasi S, Pickart CM: **Rad23 ubiquitin-associated domains (UBA) inhibit 26 S proteasome-catalyzed proteolysis by sequestering lysine 48-linked polyubiquitin chains**. *J Biol Chem* 2003, **278**(11):8951-8959.
Raasi S, Schmidtke G, Groettrup M: **The ubiquitin-like protein FAT10 forms covalent conjugates and induces apoptosis**. *J Biol Chem* 2001, **276**(38):35334-35343.
Rajagopalan KV: **Biosynthesis and processing of the molybdenum cofactors**. *Biochem Soc Trans* 1997, **25**(3):757-761.
Ramos PC, Hockendorff J, Johnson ES, Varshavsky A, Dohmen RJ: **Ump1p is required for proper maturation of the 20S proteasome and becomes its substrate upon completion of the assembly**. *Cell* 1998, **92**(4):489-499.
Rao H, Sastry A: **Recognition of specific ubiquitin conjugates is important for the proteolytic functions of the ubiquitin-associated domain proteins Dsk2 and Rad23**. *J Biol Chem* 2002, **277**(14):11691-11695.
Rechsteiner M, Hill CP: **Mobilizing the proteolytic machine: cell biological roles of proteasome activators and inhibitors**. *Trends Cell Biol* 2005, **15**(1):27-33.
Redman KL: **The smaller protein formed as a ubiquitin fusion in Drosophila is processed from ubiquitin and found on the 60S ribosomal subunit**. *Insect Biochem Mol Biol* 1994, **24**(2):191-201.
Redman KL, Rechsteiner M: **Identification of the long ubiquitin extension as ribosomal protein S27a**. *Nature* 1989, **338**(6214):438-440.
Reed JC, Doctor KS, Godzik A: **The domains of apoptosis: a genomics perspective**. *Sci STKE* 2004, **2004**(239):re9.
Regelmann J, Schule T, Josupeit FS, Horak J, Rose M, Entian KD, Thumm M, Wolf DH: **Catabolite degradation of fructose-1,6-bisphosphatase in the yeast Saccharomyces cerevisiae: a genome-wide screen identifies eight novel GID genes and indicates the existence of two degradation pathways**. *Mol Biol Cell* 2003, **14**(4):1652-1663.
Richly H, Rape M, Braun S, Rumpf S, Hoege C, Jentsch S: **A series of ubiquitin binding factors connects CDC48/p97 to substrate multiubiquitylation and proteasomal targeting**. *Cell* 2005, **120**(1):73-84.
Rost B, Liu J: **The PredictProtein server**. *Nucleic Acids Res* 2003, **31**(13):3300-3304.
Rudolph MJ, Wuebbens MM, Rajagopalan KV, Schindelin H: **Crystal structure of molybdopterin synthase and its evolutionary relationship to ubiquitin activation**. *Nat Struct Biol* 2001, **8**(1):42-46.

Ryu KS, Lee KJ, Bae SH, Kim BK, Kim KA, Choi BS: **Binding surface mapping of intra- and interdomain interactions among hHR23B, ubiquitin, and polyubiquitin binding site 2 of S5a**. *J Biol Chem* 2003, **278**(38):36621-36627.

Ryu SW, Chae SK, Kim E: **Interaction of Daxx, a Fas binding protein, with sentrin and Ubc9**. *Biochem Biophys Res Commun* 2000, **279**(1):6-10.

Ryu SW, Chae SK, Lee KJ, Kim E: **Identification and characterization of human Fas associated factor 1, hFAF1**. *Biochem Biophys Res Commun* 1999, **262**(2):388-394.

Saeki Y, Saitoh A, Toh-e A, Yokosawa H: **Ubiquitin-like proteins and Rpn10 play cooperative roles in ubiquitin-dependent proteolysis**. *Biochem Biophys Res Commun* 2002, **293**(3):986-992.

Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees**. *Mol Biol Evol* 1987, **4**(4):406-425.

Sanz P, Ludin K, Carlson M: **Sip5 interacts with both the Reg1/Glc7 protein phosphatase and the Snf1 protein kinase of Saccharomyces cerevisiae**. *Genetics* 2000, **154**(1):99-107.

Scheel H: **"Die Rolle von Homologiedomänen bei der Protein-Protein-Interaktion"**. Diploma thesis, Universität Köln; 2001.

Scheel H, Hofmann K: **No evidence for PHD fingers as ubiquitin ligases**. *Trends Cell Biol* 2003, **13**(6):285-287; author reply 287-288.

Scheel H, Hofmann K: **Prediction of a common structural scaffold for proteasome lid, COP9-signalosome and eIF3 complexes**. *BMC Bioinformatics* 2005, **6**(1):71.

Scheel H, Tomiuk S, Hofmann K: **Elucidation of ataxin-3 and ataxin-7 function by integrative bioinformatics**. *Hum Mol Genet* 2003, **12**(21):2845-2852.

Scheffner M, Nuber U, Huibregtse JM: **Protein ubiquitination involving an E1-E2-E3 enzyme ubiquitin thioester cascade**. *Nature* 1995, **373**(6509):81-83.

Scheffner M, Werness BA, Huibregtse JM, Levine AJ, Howley PM: **The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53**. *Cell* 1990, **63**(6):1129-1136.

Schild D, Brake AJ, Kiefer MC, Young D, Barr PJ: **Cloning of three human multifunctional de novo purine biosynthetic genes by functional complementation of yeast mutations**. *Proc Natl Acad Sci U S A* 1990, **87**(8):2916-2920.

Schwartz DC, Hochstrasser M: **A superfamily of protein tags: ubiquitin, SUMO and related modifiers**. *Trends Biochem Sci* 2003, **28**(6):321-328.

Schwarz SE, Matuschewski K, Liakopoulos D, Scheffner M, Jentsch S: **The ubiquitin-like proteins SMT3 and SUMO-1 are conjugated by the UBC9 E2 enzyme**. *Proc Natl Acad Sci U S A* 1998, **95**(2):560-564.

Seeler JS, Dejean A: **Nuclear and unclear functions of SUMO**. *Nat Rev Mol Cell Biol* 2003, **4**(9):690-699.

Seigneurin-Berny D, Verdel A, Curtet S, Lemercier C, Garin J, Rousseaux S, Khochbin S: **Identification of components of the murine histone deacetylase 6 complex: link between acetylation and ubiquitination signaling pathways**. *Mol Cell Biol* 2001, **21**(23):8035-8044.

Shekhtman A, Cowburn D: **A ubiquitin-interacting motif from Hrs binds to and occludes the ubiquitin surface necessary for polyubiquitination in monoubiquitinated proteins**. *Biochem Biophys Res Commun* 2002, **296**(5):1222-1227.

Shi X, Parthun MR, Jaehning JA: **The yeast EGD2 gene encodes a homologue of the alpha NAC subunit of the human nascent-polypeptide-associated complex**. *Gene* 1995, **165**(2):199-202.

Shiba T, Kawasaki M, Takatsu H, Nogi T, Matsugaki N, Igarashi N, Suzuki M, Kato R, Nakayama K, Wakatsuki S: **Molecular mechanism of membrane recruitment of GGA by ARF in lysosomal protein transport**. *Nat Struct Biol* 2003, **10**(5):386-393.

Shiba Y, Katoh Y, Shiba T, Yoshino K, Takatsu H, Kobayashi H, Shin HW, Wakatsuki S, Nakayama K: **GAT (GGA and Tom1) domain responsible for ubiquitin binding and ubiquitination**. *J Biol Chem* 2004, **279**(8):7105-7111.

Shih SC, Prag G, Francis SA, Sutanto MA, Hurley JH, Hicke L: **A ubiquitin-binding motif required for intramolecular monoubiquitylation, the CUE domain**. *Embo J* 2003, **22**(6):1273-1281.

Smith TF, Waterman MS: **Identification of common molecular subsequences**. *J Mol Biol* 1981, **147**(1):195-197.

Soares L, Seroogy C, Skrenta H, Anandasabapathy N, Lovelace P, Chung CD, Engleman E, Fathman CG: **Two isoforms of otubain 1 regulate T cell anergy via GRAIL**. *Nat Immunol* 2004, **5**(1):45-54.

Song J, Durrin LK, Wilkinson TA, Krontiris TG, Chen Y: **Identification of a SUMO-binding motif that recognizes SUMO-modified proteins**. *Proc Natl Acad Sci U S A* 2004, **101**(40):14373-14378.

Spreter T, Pech M, Beatrix B: **The Crystal Structure of Archaeal Nascent Polypeptide-associated Complex (NAC) Reveals a Unique Fold and the Presence of a Ubiquitin-associated Domain**. *J Biol Chem* 2005, **280**(16):15849-15854.

Sprinzak E, Margalit H: **Correlated sequence-signatures as markers of protein-protein interaction**. *J Mol Biol* 2001, **311**(4):681-692.

Stade K, Vogel F, Schwienhorst I, Meusser B, Volkwein C, Nentwig B, Dohmen RJ, Sommer T: **A lack of SUMO conjugation affects cNLS-dependent nuclear protein import in yeast**. *J Biol Chem* 2002, **277**(51):49554-49561.

Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF *et al*: **Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences**. *Proc Natl Acad Sci U S A* 2002, **99**(26):16899-16903.

Strohmaier H, Spruck CH, Kaiser P, Won KA, Sangfelt O, Reed SI: **Human F-box protein hCdc4 targets cyclin E for proteolysis and is mutated in a breast cancer cell line**. *Nature* 2001, **413**(6853):316-322.

Sundquist WI, Schubert HL, Kelly BN, Hill GC, Holton JM, Hill CP: **Ubiquitin recognition by the human TSG101 protein**. *Mol Cell* 2004, **13**(6):783-789.

Takahashi Y, Toh EA, Kikuchi Y: **Comparative analysis of yeast PIAS-type SUMO ligases in vivo and in vitro**. *J Biochem (Tokyo)* 2003, **133**(4):415-422.

Tan NG, Ardley HC, Scott GB, Rose SA, Markham AF, Robinson PA: **Human homologue of ariadne promotes the ubiquitylation of translation initiation factor 4E homologous protein, 4EHP**. *FEBS Lett* 2003, **554**(3):501-504.

Tanahashi N, Murakami Y, Minami Y, Shimbara N, Hendil KB, Tanaka K: **Hybrid proteasomes. Induction by interferon-gamma and contribution to ATP-dependent proteolysis**. *J Biol Chem* 2000, **275**(19):14336-14345.

Tanaka K, Suzuki T, Hattori N, Mizuno Y: **Ubiquitin, proteasome and parkin**. *Biochim Biophys Acta* 2004, **1695**(1-3):235-247.

Tatusov RL, Altschul SF, Koonin EV: **Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks**. *Proc Natl Acad Sci U S A* 1994, **91**(25):12091-12095.

Taylor SV, Kelleher NL, Kinsland C, Chiu HJ, Costello CA, Backstrom AD, McLafferty FW, Begley TP: **Thiamin biosynthesis in Escherichia coli. Identification of this thiocarboxylate as the immediate sulfur donor in the thiazole formation**. *J Biol Chem* 1998, **273**(26):16555-16560.

Terrell J, Shih S, Dunn R, Hicke L: **A function for monoubiquitination in the internalization of a G protein-coupled receptor**. *Mol Cell* 1998, **1**(2):193-202.

Thrower JS, Hoffman L, Rechsteiner M, Pickart CM: **Recognition of the polyubiquitin proteolytic signal**. *EMBO J* 2000, **19**(1):94-102.

Toth C, Coffino P: **Regulated degradation of yeast ornithine decarboxylase**. *J Biol Chem* 1999, **274**(36):25921-25926.

Tran HJ, Allen MD, Lowe J, Bycroft M: **Structure of the Jab1/MPN domain and its implications for proteasome function**. *Biochemistry* 2003, **42**(39):11460-11465.

Tsuge T, Matsui M, Wei N: **The subunit 1 of the COP9 signalosome suppresses gene expression through its N-terminal domain and incorporates into the complex through the PCI domain**. *J Mol Biol* 2001, **305**(1):1-9.

Uchida D, Hatakeyama S, Matsushima A, Han H, Ishido S, Hotta H, Kudoh J, Shimizu N, Doucas V, Nakayama KI *et al*: **AIRE functions as an E3 ubiquitin ligase**. *J Exp Med* 2004, **199**(2):167-172.

Uchihara T, Fujigasaki H, Koyano S, Nakamura A, Yagishita S, Iwabuchi K: **Non-expanded polyglutamine proteins in intranuclear inclusions of hereditary ataxias--triple-labeling immunofluorescence study**. *Acta Neuropathol (Berl)* 2001, **102**(2):149-152.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al*: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**(6770):623-627.

Unkles SE, Heck IS, Appleyard MV, Kinghorn JR: **Eukaryotic molybdopterin synthase. Biochemical and molecular studies of Aspergillus nidulans cnxG and cnxH mutants**. *J Biol Chem* 1999, **274**(27):19286-19293.

Upadhya SC, Hegde AN: **A potential proteasome-interacting motif within the ubiquitin-like domain of parkin and other proteins**. *Trends Biochem Sci* 2003, **28**(6):280-283.

Ustrell V, Hoffman L, Pratt G, Rechsteiner M: **PA200, a nuclear proteasome activator involved in DNA repair**. *EMBO J* 2002, **21**(13):3516-3525.

Valasek L, Hasek J, Nielsen KH, Hinnebusch AG: **Dual function of eIF3j/Hcr1p in processing 20 S pre-rRNA and translation initiation**. *J Biol Chem* 2001, **276**(46):43351-43360.

Valasek L, Phan L, Schoenfeld LW, Valaskova V, Hinnebusch AG: **Related eIF3 subunits TIF32 and HCR1 interact with an RNA recognition motif in PRT1 required for eIF3 integrity and ribosome binding**. *Embo J* 2001, **20**(4):891-904.

VanDemark AP, Hill CP: **Structural basis of ubiquitylation**. *Curr Opin Struct Biol* 2002, **12**(6):822-830.

Varshavsky A: **The N-end rule: functions, mysteries, uses**. *Proc Natl Acad Sci U S A* 1996, **93**(22):12142-12149.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.

Verma R, Aravind L, Oania R, McDonald WH, Yates JR, 3rd, Koonin EV, Deshaies RJ: **Role of Rpn11 metalloprotease in deubiquitination and degradation by the 26S proteasome**. *Science* 2002, **298**(5593):611-615.

Verma R, Oania R, Graumann J, Deshaies RJ: **Multiubiquitin chain receptors define a layer of substrate selectivity in the ubiquitin-proteasome system**. *Cell* 2004, **118**(1):99-110.

Wada H, Kito K, Caskey LS, Yeh ET, Kamitani T: **Cleavage of the C-terminus of NEDD8 by UCH-L3**. *Biochem Biophys Res Commun* 1998, **251**(3):688-692.

Walden H, Podgorski MS, Huang DT, Miller DW, Howard RJ, Minor DL, Jr., Holton JM, Schulman BA: **The structure of the APPBP1-UBA3-NEDD8-ATP complex reveals the basis for selective ubiquitin-like protein activation by an E1**. *Mol Cell* 2003, **12**(6):1427-1437.

Walters KJ, Kleijnen MF, Goh AM, Wagner G, Howley PM: **Structural studies of the interaction between ubiquitin family proteins and proteasome subunit S5a**. *Biochemistry* 2002, **41**(6):1767-1777.

Wang B, Alam SL, Meyer HH, Payne M, Stemmler TL, Davis DR, Sundquist WI: **Structure and ubiquitin interactions of the conserved zinc finger domain of Npl4**. *J Biol Chem* 2003, **278**(22):20225-20234.

Wang CW, Klionsky DJ: **The molecular mechanism of autophagy**. *Mol Med* 2003, **9**(3-4):65-76.

Wang L, Jiang T: **On the complexity of multiple sequence alignment**. *J Comput Biol* 1994, **1**(4):337-348.

Watkins JF, Sung P, Prakash L, Prakash S: **The Saccharomyces cerevisiae DNA repair gene RAD23 encodes a nuclear protein containing a ubiquitin-like domain required for biological function**. *Mol Cell Biol* 1993, **13**(12):7757-7765.

Wei Z, Zhang P, Zhou Z, Cheng Z, Wan M, Gong W: **Crystal Structure of Human eIF3k, the First Structure of eIF3 Subunits**. *J Biol Chem* 2004, **279**(33):34983-34990.

Weissman AM: **Themes and variations on ubiquitylation**. *Nat Rev Mol Cell Biol* 2001, **2**(3):169-178.

Wertz IE, O'Rourke KM, Zhou H, Eby M, Aravind L, Seshagiri S, Wu P, Wiesmann C, Baker R, Boone DL *et al*: **De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling**. *Nature* 2004, **430**(7000):694-699.

Wilkinson CR, Dittmar GA, Ohi MD, Uetz P, Jones N, Finley D: **Ubiquitin-like Protein Hub1 Is Required for Pre-mRNA Splicing and Localization of an Essential Splicing Factor in Fission Yeast**. *Curr Biol* 2004, **14**(24):2283-2288.

Wilkinson CR, Seeger M, Hartmann-Petersen R, Stone M, Wallace M, Semple C, Gordon C: **Proteins containing the UBA domain are able to bind to multi-ubiquitin chains**. *Nat Cell Biol* 2001, **3**(10):939-943.

Willems AR, Schwab M, Tyers M: **A hitchhiker's guide to the cullin ubiquitin ligases: SCF and its kin**. *Biochim Biophys Acta* 2004, **1695**(1-3):133-170.

Willingham S, Outeiro TF, DeVit MJ, Lindquist SL, Muchowski PJ: **Yeast genes that enhance the toxicity of a mutant huntingtin fragment or alpha-synuclein**. *Science* 2003, **302**(5651):1769-1772.

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al*: **Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis**. *Science* 1999, **285**(5429):901-906.

Witt E, Zantopf D, Schmidt M, Kraft R, Kloetzel PM, Kruger E: **Characterisation of the newly identified human Ump1 homologue POMP and analysis of LMP7(beta 5i) incorporation into 20 S proteasomes**. *J Mol Biol* 2000, **301**(1):1-9.

Xie Y, Varshavsky A: **The E2-E3 interaction in the N-end rule pathway: the RING-H2 finger of E3 is required for the synthesis of multiubiquitin chain**. *EMBO J* 1999, **18**(23):6832-6844.

Xie Y, Varshavsky A: **Physical association of ubiquitin ligases and the 26S proteasome**. *Proc Natl Acad Sci U S A* 2000, **97**(6):2497-2502.

Xie Y, Varshavsky A: **RPN4 is a ligand, substrate, and transcriptional regulator of the 26S proteasome: a negative feedback circuit**. *Proc Natl Acad Sci U S A* 2001, **98**(6):3056-3061.

Xirodimas DP, Saville MK, Bourdon JC, Hay RT, Lane DP: **Mdm2-mediated NEDD8 conjugation of p53 inhibits its transcriptional activity**. *Cell* 2004, **118**(1):83-97.

Yamano T, Murata S, Shimbara N, Tanaka N, Chiba T, Tanaka K, Yui K, Udono H: **Two distinct pathways mediated by PA28 and hsp90 in major histocompatibility complex class I antigen processing**. *J Exp Med* 2002, **196**(2):185-196.

Yamazaki K, Guo L, Sugahara K, Zhang C, Enzan H, Nakabeppu Y, Kitajima S, Aso T: **Identification and biochemical characterization of a novel transcription elongation factor, Elongin A3**. *J Biol Chem* 2002, **277**(29):26444-26451.

Yao T, Cohen RE: **A cryptic protease couples deubiquitination and degradation by the proteasome**. *Nature* 2002, **419**(6905):403-407.

Yaseen NR, Blobel G: **Two distinct classes of Ran-binding sites on the nucleoporin Nup-358**. *Proc Natl Acad Sci U S A* 1999, **96**(10):5516-5521.

Yashiroda H, Tanaka K: **Hub1 is an essential ubiquitin-like protein without functioning as a typical modifier in fission yeast**. *Genes Cells* 2004, **9**(12):1189-1197.

Yeh ET, Gong L, Kamitani T: **Ubiquitin-like proteins: new wines in new bottles**. *Gene* 2000, **248**(1-2):1-14.

Young P, Deveraux Q, Beal RE, Pickart CM, Rechsteiner M: **Characterization of two polyubiquitin binding sites in the 26 S protease subunit 5a**. *J Biol Chem* 1998, **273**(10):5461-5467.

Yuan W, Krug RM: **Influenza B virus NS1 protein inhibits conjugation of the interferon (IFN)-induced ubiquitin-like ISG15 protein**. *EMBO J* 2001, **20**(3):362-371.

Yuan X, Simpson P, McKeown C, Kondo H, Uchiyama K, Wallis R, Dreveny I, Keetch C, Zhang X, Robinson C *et al*: **Structure, dynamics and interactions of p47, a major adaptor of the AAA ATPase, p97**. *EMBO J* 2004, **23**(7):1463-1473.

Zachariae W, Shevchenko A, Andrews PD, Ciosk R, Galova M, Stark MJ, Mann M, Nasmyth K: **Mass spectrometric analysis of the anaphase-promoting complex from yeast: identification of a subunit related to cullins**. *Science* 1998, **279**(5354):1216-1219.

Zaiss DM, Standera S, Holzhutter H, Kloetzel P, Sijts AJ: **The proteasome inhibitor PI31 competes with PA28 for binding to 20S proteasomes**. *FEBS Lett* 1999, **457**(3):333-338.

Zaiss DM, Standera S, Kloetzel PM, Sijts AJ: **PI31 is a modulator of proteasome formation and antigen processing**. *Proc Natl Acad Sci U S A* 2002, **99**(22):14344-14349.

Zeth K, Ravelli RB, Paal K, Cusack S, Bukau B, Dougan DA: **Structural analysis of the adaptor protein ClpS in complex with the N-terminal domain of ClpA**. *Nat Struct Biol* 2002, **9**(12):906-911.

Zhang SQ, Kovalenko A, Cantarella G, Wallach D: **Recruitment of the IKK signalosome to the p55 TNF receptor: RIP and A20 bind to NEMO (IKKgamma) upon receptor stimulation**. *Immunity* 2000, **12**(3):301-311.

Zhao C, Beaudenon SL, Kelley ML, Waddell MB, Yuan W, Schulman BA, Huibregtse JM, Krug RM: **The UbcH8 ubiquitin E2 enzyme is also the E2 enzyme for ISG15, an IFN-alpha/beta-induced ubiquitin-like protein**. *Proc Natl Acad Sci U S A* 2004, **101**(20):7578-7582.

Zhou C, Wee S, Rhee E, Naumann M, Dubiel W, Wolf DA: **Fission yeast COP9/signalosome suppresses cullin activity through recruitment of the deubiquitylating enzyme Ubp12p**. *Mol Cell* 2003, **11**(4):927-938.

Zhu C, Karplus K, Grate L, Coffino P: **A homolog of mammalian antizyme is present in fission yeast Schizosaccharomyces pombe but not detected in budding yeast Saccharomyces cerevisiae**. *Bioinformatics* 2000, **16**(5):478-481.

6 Appendix

Table 6-1 Human RING finger proteins. Continued on the next page.

Human RING finger proteins			
Gene name	**Uniprot number**	**Gene name**	**Uniprot number**
AMFR	Q9UKV5	MARCH1	Q8TCQ1
ANAPC11	Q9NYG5	MARCH2/MARCH-II	Q9P0N8
BARD1	Q99728	MARCH3/MARCH-III	Q86UD3
BFAR/BAR	Q9NZS9	MARCH5/RNF153	Q9NX47
BIRC2/cIAP1	Q13490	MARCH6/KIAA0597	O60337
BIRC3/cIAP2	Q13489	MARCH7/AXOT	Q9H992
BIRC4/XIAP	P98170	MARCH8/MIR	Q8TC72
BIRC7/Livin	Q96CA5	MARCH9/MARCH-IX	Q86VN5
BIRC8/ILP2	Q96P09	MDM2	Q00987
BRAP	Q7Z569	MDM4	O15151
BRCA1	P38398	MGC4734	Q96D59
C13orf7	Q5W0B1	MGRN1	Q86W76
C16orf28/FLJ12623	Q9H9P5	MIB1/MIB	Q86YT6
C17orf29	Q63HN8	MID1	O15344
C18orf23/FLJ45559	Q6ZSG1	MID2	Q9UJV3
CBL	P22681	MKRN1	Q9UHC7
CBLB	Q13191	MKRN2	Q9H000
CBLC	Q9ULV8	MKRN3	Q13064
CBLL1	Q8TAJ4	MNAB/MASNAB	Q9HBD2
CGRRF1	Q99675	MNAT1	P51948
CHFR	Q96EP1	MYCBP2	Q6PIB6
CNOT4	O95628	MYLIP	Q8WY64
DKFZp434E1818	ENSP00000343122	NEURL	O76050
DKFZp547C195	Q6P2E0	NFX1	Q12986
DKFZp761H1710	Q9H0X6	NHLRC1/Malin	Q6VVB1
DTX1	Q86Y01	NSMCE1/NSE1	Q8WV22
DTX2	Q86UW9	PCGF1/NSPC1	Q9BSM1
DTX3	Q8N9I9	PCGF2/RNF110	P35227
DTX3L/BBAP	Q8TDB6	PCGF4/BMI1	P35226
DTX4/KIAA0937	Q9Y2E6	PCGF5	Q86SE9
DZIP3	Q86Y13	PCGF6/hMBLR	Q9BYE7
ENSP00000280266	ENSP00000280266	PDZRN3/KIAA1095	Q9UPQ7
ENSP00000344026	ENSP00000344026	PEX10	O60683
ENSP00000348371	ENSP00000348371	PEX12	O00623
FLJ10520	Q5XKR3	PHF7	Q9NSX7
FLJ12270/KIAA1923	Q96PW5	PJA1	Q8NG27
FLJ12875	Q969V5	PJA2	Q8N1G5
FLJ16581	Q6ZWI9	PML	P29590
FLJ20225	Q9NXI6	PXMP3	P28328
FLJ20315/URCC	Q65ZA4	RAD18	Q9NS91
FLJ23749	Q8TEA0	RAG1	P15918
FLJ31951	Q8IVP7	RAPSN	Q13702
FLJ35757	Q8NA82	RBBP6	Q7Z6E9
FLJ36180	Q8N9V2	RBX1	P62877
FLJ38628	Q96GF1	RCHY1	Q96PM5
FLJ45273	Q6ZSR4	RFFL	Q8TBY7
FLJ46380	Q6ZRF8	RFP	P14373
GENSCAN00000024511	GENSCAN00000024511H	RFP2	O60858
SHPRH	Q8IWQ9	RFPL1	O75677
HOZFP	Q86VG1	RFPL2	O75678
KIAA0804	O94896	RFPL3	O75679
KIAA1333/FLJ20333	Q9NXC0	RFWD2/COP1	Q8NHY2
KIAA1404	Q9P2E3	RING1	Q06587
KIAA1542	Q9P1Y6	RKHD1	Q86XN8
KIAA1972	Q96DX4	RKHD2	Q5U5Q3
KIAA1991	Q8NCN4	RKHD3/KIAA2009	Q8IVG2
LNX	Q8TBB1	RNF10/RIE2	Q9ULW4
LNX2	Q8N448	RNF103	O00237
LOC149603	Q6PJR0	RNF11	Q9Y3C5
LOC285498	Q8IY99	RNF111	Q6P9A4
LOC493829	Q8N4X6	RNF12	Q9NVW2
LOC51136/FLJ25783	Q8N7D0	RNF121	Q96DB4
LOC51255	Q9P0P0	RNF122/FLJ12526	Q9H9V4
LRSAM1/TAL	Q6UWE0	RNF123/KPC1	Q5XPI4
M96/MTF2	Q9Y483	RNF125	Q96EQ8
MAP3K1	O13233	RNF126	Q9BV68

Human RING finger proteins			
Gene name	**Uniprot number**	**Gene name**	**Uniprot number**
RNF127/FLJ34458	Q8NB00	TRIM28	Q13263
RNF128/GRAIL	Q96RF3	TRIM3	O75382
RNF13	O43567	TRIM31	Q9BZY9
RNF130	Q86XS8	TRIM32	Q13049
RNF133	Q8WVZ7	TRIM33	Q9UPN9
RNF135/MGC13061	Q8IUD6	TRIM34	Q9BYJ4
RNF138	Q8WVD3	TRIM35	Q9UPQ4
RNF139/TRC8	O75485	TRIM36	Q9NQ86
RNF141	Q8WVD5	TRIM37	O94972
RNF146/Dactylidin	Q9NTX7	TRIM38	O00635
RNF148	Q8N308	TRIM39/RNF23	Q9HCM9
RNF149	Q8NC42	TRIM4	Q9C037
RNF150/KIAA1214	Q9ULK6	TRIM40/RNF35	Q6P9F5
RNF151	Q8NHS5	TRIM41	Q8WV44
RNF152	Q8N8N0	TRIM42	Q8IWZ5
RNF157/KIAA1917	Q96PX1	TRIM43	Q96BQ3
RNF166	Q96A37	TRIM45	Q9H8W5
RNF167	Q9H6Y7	TRIM46	Q7Z4K8
RNF168/FLJ35794	Q8IYW5	TRIM47	Q96LD4
RNF170	Q86YC0	TRIM48	Q8IWZ4
RNF175/LOC285533	Q8N4F7	TRIM49/RNF18	Q9NS80
RNF180	Q86T96	TRIM5	Q9C035
RNF182/MGC33993	Q8N6D2	TRIM50A	Q86XT4
RNF2/DING	Q99496	TRIM50B	Q86UV7
RNF20	Q5VTR2	TRIM50C	Q86UV6
RNF24	Q9Y225	TRIM52	Q96A61
RNF25	Q96BH1	TRIM54/RNF30	Q9BYV2
RNF26	Q9BY78	TRIM55/RNF29	Q9BYV6
RNF32	Q6FIB3	TRIM56	Q9BRZ2
RNF34	Q969K3	TRIM58/BIA2	Q8NG06
RNF36	Q86WT6	TRIM59/TSBF1	Q8IWR1
RNF38	Q9H0F5	TRIM6	Q9C030
RNF39/HCGV	Q96QB5	TRIM60/FLJ35882	Q8NA35
RNF3A	O15262	TRIM61	Q5EBN2
RNF4	P78317	TRIM62	Q9BVG3
RNF40/KIAA0661	O75150	TRIM63/RNF28	Q969Q1
RNF41	O75598	TRIM65	Q6PJ69
RNF44	Q7L0R7	TRIM67/TNL	Q7Z4K7
RNF5/HsRma1	Q99942	TRIM68	Q6AZZ1
RNF6	Q9Y252	TRIM7	Q9C029
RNF7/ROC2	Q9UBF6	TRIM8	Q9BZR9
RNF8	O76064	TRIM9	Q9C026
RP11-307C12.10	Q5T197	TRIP/TRAIP	Q9BWF2
RP4-678E16.1	Q5VTB9	TTC3	P53804
RP5-1198E17.5	Q5TC82	UBOX5/RNF37	O94941
SH3MD2	Q7Z6J0	UBR1	Q8IWV7
SH3RF2/FLJ23654	Q8TEC5	UBR2/UBR1L2	Q8IWV8
SIAH1	Q8IUQ4	UHRF1/NP95	Q96T88
SIAH2	O43255	UHRF2/NIRF	Q96PU4
SMARCA3/HIP116	Q14527	VPS11	Q9H270
SYVN1/HRD1	Q8N6E8	VPS18	Q9P253
TOPORS	Q9UNR9	VPS41	P49754
TRAF2	Q12933	ZFPL1	O95159
TRAF3	Q13114	ZNF179	Q9ULX5
TRAF4	Q9BUZ4	ZNF183	O15541
TRAF5	O00463	ZNF183L1	Q8IZP6
TRAF6	Q9Y4K3	ZNF294	O94822
TRAF7	Q6Q0C0	ZNF313	Q9Y508
TRIM10	Q9UDY6	ZNF364	Q9Y4L5
TRIM11	Q96F44	ZNF598	Q86UK7
TRIM15	Q9C019	ZNF645	Q6DJY9
TRIM17	Q9Y577	ZNF650/UBR1L1	Q6ZT12
TRIM2	Q9C040	ZNRF1	Q8ND25
TRIM21	P19474	ZNRF2	Q8NHG8
TRIM22	Q8IYM9	ZNRF3/KIAA1133	Q9ULT6
TRIM23	P36406	ZNRF4/LOC148066	Q8WWF5
TRIM24/TIF1	O15164	ZSWIM2	Q8NEG5
TRIM25	Q14258	ZZANK1/Skeletrophin	Q8NI59
TRIM26	Q12899	*MKRN4*	*Q13434*

Table 6-2 Yeast RING finger proteins.

Yeast RING finger proteins			
Gene name	**ORF**	**Gene name**	**ORF**
APC11	YDL008W	RAD18	YCR066W
ASI1	YMR119W	RAD5	YLR032W
ASR1	YPR093C	RIS1	YOR191W
BRE1	YDL074C	SAN1	YDR143C
CWC24	YLR323C	SLX8	YER116C
DMA1	YHR115C	SSM4	YIL030C
DMA2	YNL116W	STE5	YDR103W
FAP1	YNL023C	TFB3	YDR460W
FAR1	YJL157C	TUL1	YKL034W
HRD1	YOL013C	UBR1	YGR184C
HRT1	YOL133W	UBR2	YLR024C
MAG2	YLR427W	VPS8	YAL002W
MOT2	YER068W	YBR062C	YBR062C
PEP3	YLR148W	YDR128W	YDR128W
PEP5	YMR231W	YDR266C	YDR266C
PEX10	YDR265W	YHL010C	YHL010C
PEX12	YMR026C	YLR247C	YLR247C
PIB1	YDR313C	YMR247C	YMR247C
PSH1	YOL054W	YOL138C	YOL138C
RAD16	YBR114W		

Table 6-3 Orthology assignments for yeast and human RING finger proteins.

Orthology assignments for yeast and human RING finger proteins	
Yeast	**Human**
APC11	ANAPC11
BRE1	RNF20 RNF40/KIAA0661
CWC24	ZNF183 ZNF183L1
FAP1	NFX1
DMA1	CHFR
DMA2	RNF8
HRD1	SYVN1/HRD1
HRT1	RBX1
MAG2	RIE2
MOT2	CNOT4
PEP3	VPS18
PEP5	VPS11
PEX10	PEX10
PEX12	PEX12
RAD18	RAD18
RAD5 RAD16 RIS1	SMARCA3/HIP116
SSM4	MARCH6/KIAA0597
TFB3	MNAT1
VPS8	KIAA0804
YDR128W	FLJ12270/KIAA1923
YDR266C	ZNF598
YHL010C	BRAP
YMR247C	ZNF294
UBR1 UBR2	UBR1 UBR2 UBR1L1/FLJ45053

Table 6-4 Human F-Box proteins.

Human F-Box proteins					
Gene name	Uniprot number	additional motifs	Gene name	Uniprot number	additional motifs
FBXL1/SKP2	Q13309	LRR	FBXO17/FBXO26	Q96EF6	FBA
FBXL2	Q9UKC9	LRR	FBXO18	Q8NFZ0	others/none
FBXL3	Q9UKT7	LRR	FBXO21	O94952	others/none
FBXL4	Q9UKA2	LRR	FBXO22	Q8NEZ5	others/none
FBXL5	Q9UKA1	LRR	FBXO24	O75426	GRF_RCC
FBXL6	Q8N531	LRR	FBXO25	Q8TCJ0	others/none
FBXL7	Q9UJT9	LRR	FBXO27	Q8NI29	FBA
FBXL8	Q96CD0	LRR	FBXO28	Q9NVF7	others/none
FBXL9/LRRC29	Q8WV35	LRR	FBXO30	Q8TB52	ZF_TRAF
FBXL10	Q8NHM5	LRR	FBXO31/FBXO14	Q5XUX0	others/none
FBXL11	Q9Y2K7	LRR	FBXO32	Q969P5	others/none
FBXL12	Q9NXK8	LRR	FBXO33	Q7Z6M2	others/none
FBXL13	Q8NEE6	LRR	FBXO34	Q9NWN3	others/none
FBXL14	Q8N1E6	LRR	FBXO36	Q8NEA4	others/none
FBXL15/FBXO37	Q9H469	LRR	FBXO38	Q6PIJ6	others/none
FBXL16	Q8N461	LRR	FBXO39	Q8N4B4	others/none
FBXL17/FBXO13	Q9UF56	LRR	FBXO40	Q9UH90	ZF_TRAF
FBXL18	Q96ME1	LRR	FBXO41	Q8TF61	others/none
FBXL19	Q6PCT2	LRR	FBXO42	Q6P3S6	KELCH
FBXL20	Q96IG2	LRR	FBXO43	ENSP00000322600	others/none
FBXL21	Q9UKT6	LRR	FBXO44	Q9H4M3	FBA
FBXL22	Q6P050	LRR	FBXO45	ENSP00000310332	others/none
FBXO1/CCNF	P41002	LRR	FBXO46	Q6PJ61	others/none
FBXO2	Q9UK22	FBA	FBXW1/BTRC	Q9Y297	WD40
FBXO3	Q9UK99	others/none	FBXW2	Q9UKT8	WD40
FBXO4	Q9UKT5	others/none	FBXW3	Q9UKB7	WD40
FBXO5	Q9UKT4	others/none	FBXW4/SHFM3	P57775	WD40
FBXO6	Q9NRD1	FBA	FBXW5	Q969U6	WD40
FBXO7	Q9Y3I1	UbL, PI31_NTERM	FBXW7/FBXW6	Q969H0	WD40
FBXO8	Q9NRD0	SEC7	FBXW8/FBXO29	Q8N3Y1	WD40
FBXO9	Q9UK97	others/none	FBXW9	Q5XUX1	WD40
FBXO10	Q9UK96	others/none	FBXW10	Q5XX13	WD40
FBXO11	Q86XK2	ZF_UBR1	FBXW11	Q9UKB1	WD40
FBXO15	Q8NCQ5	others/none	FBXW12/FBXO35	Q6X9E4	others/none
FBXO16	O8IX29	others/none			

Table 6-5 Yeast F-Box proteins.

Yeast F-Box proteins		
Gene name	ORF	additional motifs
YBR280C	YBR280C	GRF_RCC
AMN1	YBR158W	LRR
DIA2	YOR080W	LRR
GRR1	YJR090C	LRR
YLR352W	YLR352W	LRR
UFO1	YML088W	3xUIM
CDC4	YFL009W	WD40
MET30	YIL046W	WD40
COS111	YBR203W	others/none
HRT3	YLR097C	others/none
MDM30	YLR368W	others/none
RCY1	YJL204C	others/none
YDR131C	YDR131C	others/none
YDR219C	YDR219C	others/none
YDR306C	YDR306C	others/none
YJL149W	YJL149W	others/none
YLR224W	YLR224W	others/none
YMR258C	YMR258C	others/none
SKP2	YNL311C	others/none
CTF13	YMR094W	others/none

Table 6-6 Human SOCS proteins.

Human SOCS proteins		
Gene name	**ORF**	**additional motifs**
TULP4	Q9NRJ4	WD40
WSB1	Q9Y6I7	WD40
WSB2	Q9NYS7	WD40
ASB1	Q9Y576	ANKYRIN
ASB2	Q96Q27	ANKYRIN
ASB3	Q9Y575	ANKYRIN
ASB4	Q9Y574	ANKYRIN
ASB5	Q8WWX0	ANKYRIN
ASB6	Q9NWX5	ANKYRIN
ASB7	Q9H672	ANKYRIN
ASB8	Q9H765	ANKYRIN
ASB9	Q96DX5	ANKYRIN
ASB10	Q8WXI3	ANKYRIN
ASB11	Q8WXH4	ANKYRIN
ASB12	Q8WXK4	ANKYRIN
ASB13	Q8WXK3	ANKYRIN
ASB14	Q8WXK2	ANKYRIN
ASB15	Q8WXK1	ANKYRIN
ASB16	Q96NS5	ANKYRIN
ASB17	Q8WXJ9	ANKYRIN
RAB40A	Q8WXH6	GTPASE_RAB
RAB40B	Q12829	GTPASE_RAB
RAB40C	Q96S21	GTPASE_RAB
SOCS1	O15524	SH2
SOCS2	O14508	SH2
SOCS3	O14543	SH2
SOCS4	Q8WXH5	SH2
SOCS5	O75159	SH2
SOCS6	O14544	SH2
SOCS7	O14512	SH2
CISH	Q9NSE2	SH2
SSB1	Q96BD6	SPRY
SSB3	Q96IE6	SPRY
SSB4	Q96A44	SPRY
GRCC9/SSB2	Q99619	SPRY
LOC196394	Q8IY45	LRR
TCEB3	Q14241	TFII2_ELONGIN
TCEB3B	Q8IYF1	TFII2_ELONGIN
TCEB3C	Q8NG57	TFII2_ELONGIN
NEURL2	Q9BR09	others/none
VHL	P40337	others/none

Table 6-7 Yeast SOCS proteins.

Yeast SOCS proteins		
Gene name	**ORF**	**additional motifs**
ELA1	YNL230C	others/none
RAD7	YJR052W	LRR

Table 6-8 Human BTB proteins.

Human BTB proteins					
Gene name	ORF	additional motifs	Gene name	ORF	additional motifs
BKLHD5/KIAA1900	Q96NJ5	KELCH	TZFP/FAZF	Q9Y2Y4	ZF
BTBD5	Q9NXS3	KELCH	ZBTB1	Q9Y2K1	ZF
C16orf44	Q8N4N3	KELCH	ZBTB2	Q8N680	ZF
CCIN	Q13939	KELCH	ZBTB3	Q9H5J0	ZF
DRE1	Q6TFL4	KELCH	ZBTB4	Q9P1Z0	ZF
ENC1	O14682	KELCH	ZBTB5	O15062	ZF
ENC2/DKFZp434K111	Q9H0H3	KELCH	ZNF482/ZBTB6	Q15916	ZF
FLJ11078	Q8TAP0	KELCH	ZBTB7A	O95365	ZF
FLJ34960	Q8N239	KELCH	ZFP67/ZBTB7B/ZBTB15	O15156	ZF
FLJ43374	Q6ZUS1	KELCH	ZBTB8a	Q8NAP8	ZF
IPP	Q9Y573	KELCH	ZBTB8b	Q96BR9	ZF
IVNS1ABP	Q9Y6Y0	KELCH	ZBTB9	Q96C00	ZF
KBTBD2	Q8IY47	KELCH	ZBTB10	Q96DT7	ZF
KBTBD3	Q8NAB2	KELCH	ZBTB11	O95625	ZF
KBTBD4	Q9NVX7	KELCH	ZBTB12	Q9Y330	ZF
KBTBD5	Q86SI1	KELCH	ZFP161/ZBTB14	O43829	ZF
KBTBD6	Q86V97	KELCH	ZBTB16	Q05516	ZF
KBTBD7	Q8WVZ9	KELCH	ZBTB17	Q13105	ZF
KBTBD9	Q96CT2	KELCH	ZNF238/ZBTB18	Q99592	ZF
KBTBD10	O60662	KELCH	ZNF278/ZBTB19	Q9HBE1	ZF
KELCHL	Q96B68	KELCH	ZBTB20	Q9HC78	ZF
KIAA0711	O94819	KELCH	ZNF295/ZBTB21	Q9ULJ3	ZF
KIAA1340	Q9P2K6	KELCH	ZNF297/ZBTB22A	O15209	ZF
KLHL1	Q9NR64	KELCH	ZNF297B/ZBTB22B	O43298	ZF
KLHL2	O95198	KELCH	ZNF336/ZBTB23	Q9H116	ZF
KLHL3	Q9UH77	KELCH	ZBTB24	O43167	ZF
KLHL4	Q9C0H6	KELCH	ZNF46/ZBTB25	P24278	ZF
KLHL5	Q96PQ7	KELCH	ZBTB26	Q9HCK0	ZF
KLHL6	Q8WZ60	KELCH	BCL6/ZBTB27	P41182	ZF
KLHL7	Q8IXQ5	KELCH	BCL6B/ZBTB28	Q8N143	ZF
KLHL8	Q9P2G9	KELCH	HIC1/ZBTB29	Q14526	ZF
KLHL9	Q9P2J3	KELCH	HIC2/ZBTB30	Q96JB3	ZF
KLHL10	Q6JEL2	KELCH	MYNN/ZBTB31	Q86Z12	ZF
KLHL11	Q9NVR0	KELCH	ZBTB33/kaiso	Q86T24	ZF
KLHL12	Q9HBX5	KELCH	ZBTB34	Q8NCN2	ZF
KLHL13	Q9P2N7	KELCH	ZBTB37	Q5TC79	ZF
KLHL14	Q9P2G3	KELCH	ZNF131	P52739	ZF
KLHL15	Q96M94	KELCH	ZNF499	Q96K62	ZF
GAN/KLHL16	Q9H2C0	KELCH	ZNF509/FLJ45653	Q6ZSB9	ZF
KLHL17	Q6TDP4	KELCH	ZNF651/FLJ45122	Q6ZSY6	ZF
KLHL18	O94889	KELCH	RHOBTB1	O94844	GTPASE_RHO
KEAP1/KLHL19	Q14145	KELCH	RHOBTB2	Q9BYZ6	GTPASE_RHO
KLEIP/KLHL20	Q9Y2M5	KELCH	RHOBTB3	O94955	others/none
KLHL21	Q9UJP4	KELCH	SPOP	O43791	MATH
LZTR1	Q8N653	KELCH	LOC339745	Q6IQ16	MATH
MGC2610	Q8NBE8	KELCH	BACH1	O14867	others/none
OTTHUMP00000016633	Q9H511	KELCH	BACH2	Q9BYV9	others/none
TA-KRP/KIAA1842	Q96JI5	KELCH	BTBD1	Q9H0C5	others/none
BTBD11/FLJ42845	Q6ZV99	ANKYRIN	BTBD12	Q8IY92	others/none
ABTB2	Q8N961	ANKYRIN	BTBD14A	Q96BF6	others/none
ABTB1	Q969K4	ANKYRIN	BTBD14B	Q96RE7	others/none
IBTK	Q9P2D0	ANKYRIN/GRF_RCC	BTBD2	Q9BX70	others/none
ANKFY1	Q9P2R3	ANKYRIN	BTBD3	Q9Y2F9	others/none
CHC1L	O95199	GRF_RCC	BTBD6	Q96KE9	others/none
RCBTB1	Q8NDN9	GRF_RCC	BTBD8	Q5XKL5	others/none
APM-1	O73453	ZF	BTBD9	Q96Q07	others/none
BTBD4	Q86UZ6	ZF	C10orf87	Q96LN0	others/none
FLJ35036	Q8NAP3	ZF	GMCL1/GCL	Q96IK5	others/none
FRBZ1	Q8IZ99	ZF	GMCL1L	Q8NEA9	others/none
HKR3	P10074	ZF	BTBD7/KIAA1525	Q9P203	others/none
HSPC063	Q8NCP5	ZF	LGALS3BP	Q08380	others/none
KIAA0352	O15060	ZF	GENSCAN00000050486H	GENSCAN00000050486H	others/none
KIAA0478	Q9NUA8	ZF	LOC149478	GENSCAN00000058813H	others/none

Table 6-9 Yeast BTB proteins.

Yeast BTB proteins		
Gene name	ORF	additional motifs
YIL001W	YIL001W	ANKYRIN
MDS3	YGL197W	KELCH
PMD1	YER132C	KELCH
YDR132C	YDR132C	others/none
WHI2	YOR043W	others/none
YLR108C	YLR108C	others/none

7 Acknowledgments

First I would like to thank Dr. Kay Hofmann, who has supervised this project primarily, for his patient support over the last four years and his great scientific knowledge that helped to guide and interpret my work. Especially, I want to thank him for let me work on a topic of outstanding interest and to teach me the bioinformatical skills that so often gave deep insights into protein evolution during the preparation of this work. I hope to keep up and deepen these skills to further explore evolutionary connections in order to contribute to the research of Dr. Kay Hofmann and the scientific community in general. In addition and perhaps most importantly, I want to thank him for teaching me how to think critically about my research and how to present my work in a scientific manner. He has also provided the computational infrastructure, a wealth of biological databases and programming advices. I also want to thank him for that. Kay has occupied a central position in this work in every respect.

I would also like to thank Memorec Biotec GmbH, Cologne, represented by Dr. Boris Stoffel for giving me the opportunity to carry out my doctoral thesis. In this respect, I also thank Stefan Miltenyi very much for assisting Memorec in hard times.

I am sincerely grateful to Prof. Dr. Jürgen Dohmen for academic supervision and his helpful advices. Without his interest and commitment, this work would never have been realized. In addition, I thank him and Palanimurugan Rangasani for experimental workout and successful publication of the yeast Antizyme project.

I would like to thank Bettina Hetfeld and Prof. Dr. W. Dubiel for sharing their experimental data with me and for a successful collaboration on human USP15.

I am grateful to Prof. Dr. Jürgen Dohmen, Kristina Uzunova and Marion Schnellhardt for providing me with data on SUMO interactors and for helpful discussion.

I would like to thank the remaining members of my committee, Prof. Dr. R. Krämer, Prof. Dr. T. Langer and the assessor, Dr. M. Cramer.

Special thanks go to Dr. Stefan Tomiuk for encouraging and driving me during the last four years. Also thanks for integrating me into Memorec projects and teaching me perfect microarray probe design. Special thanks also to the former member of our team, Dr. Frank Schacherer, for technical support and programming style discussions. He is right: proper documentation is the key for programming success. I think, we three and Kay, of course, had a great time the last years - thank you for that!

Bei allen Mitarbeitern der Memorec Biotec GmbH bedanke ich mich für ihre stete Unterstützung und große Hilfsbereitschaft.

Meiner Familie, die durch ihre menschliche Unterstützung zum Gelingen dieser Arbeit beigetragen haben, danke ich sehr herzlich.

Mein besonderer Dank gilt meinen Eltern, die mich während all der Jahre begleitet und mir stets großen Rückhalt gegeben haben.

Hartmut Scheel

8 Eidesstattliche Erklärung

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Jürgen Dohmen betreut worden.

Köln, 16. Mai 2005

Hartmut Scheel

Teilpublikationen

1. Scheel, H., S. Tomiuk and K. Hofmann (2002). "A common protein interaction domain links two recently identified epilepsy genes." Hum Mol Genet **11**(15): 1757-62.
2. Scheel, H. and K. Hofmann (2003). "A novel interaction motif, SARAH, connects three classes of tumor suppressor." Curr Biol **13**(23): R899-900.
3. Scheel, H. and K. Hofmann (2003). "No evidence for PHD fingers as ubiquitin ligases." Trends Cell Biol **13**(6): 285-7; author reply 287-8.
4. Scheel, H., S. Tomiuk and K. Hofmann (2003). "Elucidation of ataxin-3 and ataxin-7 function by integrative bioinformatics." Hum Mol Genet **12**(21): 2845-52.
5. Palanimurugan, R., H. Scheel, K. Hofmann and R. J. Dohmen (2004). "Polyamines regulate their synthesis by inducing expression and blocking degradation of ODC antizyme." EMBO J **23**(24): 4857-67. Epub 2004 Nov 11.
6. Scheel, H. and K. Hofmann (2005). "Prediction of a common structural scaffold for proteasome lid, COP9-signalosome and eIF3 complexes." BMC Bioinformatics **6**(1): 71.
7. Hetfeld, B. K. J., A. Helfrich, B. Kapelari, H. Scheel, K. Hofmann, A. Guterman, M. Glickman, W. Dubiel, R. Schade and P. Kloetzel (2005). "The COP9 signalosome associated deubiquitinating enzyme USP15 exhibits a Zinc-finger that is essential for rescuing the ubiquitin ligase Rbx1 from degradation." Curr Biol. 2005 Jul 12;15(13):1217-21.