Université de Montréal

Structural bioinformatics analysis of the family of human ubiquitin-specific proteases

par Xiao Zhu

Département de biochimie
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences en bio-informatique

Décembre 2007

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Structural bioinformatics analysis of the family of human ubiquitin-specific proteases

présenté par :

Xiao Zhu

a été évalué par un jury composé des personnes suivantes :

Jurgen Sygusch
président-rapporteur

Traian Sulea
directeur de recherche

Robert Ménard
codirecteur

Simon Wing
membre du jury

Résumé

Les protéases spécifiques à l'ubiquitine (« ubiquitin-specific proteases » ou USPs) représentent un sous-ensemble important d'enzymes de dé-ubiquitination humaines (DUBs) qui catalysent la libération d'ubiquitine des protéines ubiquitinées. L'annotation fonctionnelle de ces enzymes est limitée aux méthodes de comparaison par séquences par l'absence de données structurelles. Ainsi, la fonction des régions longues en amont et en aval du domaine catalytique ainsi que de celles encastrées à l'intérieur demeure inconnue. Dans l'article *High Incidence of Ubiquitin-like Domains in Human Ubiquitin-Specific Proteases* ci-présenté, nous démontrons la présence de nouvelles régions structurées et fonctionnelles chez les USPs humaines. Afin d'élargir notre compréhension sur la fonction de ces enzymes, nous avons appliqué la méthode de prédiction de repliement par consensus aux régions non-annotées des USPs. Notre prédiction à haute fiabilité suggère que le domaine apparenté à l'ubiquitine (« ubiquitin-like » ou UBL) constitue le domaine le plus fréquent dans la famille des USPs humaines. La présence des UBLs en copies multiples ainsi qu'insérés au sein du domaine catalytique de certaines USPs démontrent d'un haut degré de complexité structurale. Ces données complémentent nos connaissances actuelles sur l'organisation structurelle et fonctionnelle de l'ensemble des DUBs. Nos résultats suggèrent que la présence des UBLs est quasiment aussi fréquente que le domaine catalytique. Nous proposons diverses fonctions possibles pour des nouveaux domaines UBLs découverts chez les USPs humaines, telles que l'association avec le protéasome, le recrutement spécifique de substrats, ainsi que la distribution intracellulaire.

Mots clés : dé-ubiquitination, prédiction de repliement, protéasome, ubiquitine, UBL, USP.

Abstract

The family of ubiquitin-specific proteases (USPs) is the major member of the human deubiquitinating enzymes (DUBs) superfamily that specifically cleaves ubiquitin from ubiquitin-conjugated substrates. Current functional annotations of USPs is limited to sequence comparison methods and to the minimal availability of crystallographic data. Large regions both within and flanking the catalytic core which may explain substrate specificity and subcellular localization remain poorly defined. In the article presented, *High Incidence of Ubiquitin-like Domains in Human Ubiquitin-Specific Proteases*, we provide evidence of the presence of novel structural features and domains within the human USPs. Our methods utilize consensus protein fold recognition techniques on stretches of un-annotated regions in the USPs in order to expand our current understanding of the functional role of these enzymes. Among other interesting findings, we have discovered a high occurrence of reliably predicted ubiquitin-like (UBL) folds situated at both N- and C-terminal regions as well as embedded within the catalytic core of human USPs. The marked presence of multiple UBL domains as well as those integrated within the catalytic core present a new dimension in the structural complexity of these enzymes. Our results also suggest that the occurrence of UBL domains in human USPs is a close second to the characteristic catalytic core. Several propositions for the functional implications including proteasome binding and substrate recognition of these UBLs are discussed.

Keywords: deubiquitination, consensus fold recognition, proteasome, ubiquitin, UBL, USP

Table of Contents

Acknowledgements

I'd like to thank my advisors Dr. Traian Sulea and Dr. Robert Ménard who, with their rich and diverse expertise, have given me the opportunity to conduct this exciting research project. Throughout my journey at Biotechnology Research Institute (BRI), they gave academic guidance and mentorship that helped me to succeed. I was able to make critical decisions for my project and acquire independency in research. This valuable experience and many lessons learned throughout my stay have thoroughly prepared me for my scientific career.

In addition, I would like to thank Dr. Enrico Purisima for allowing me to take part in the Computational Chemistry group at BRI and for securing funds to carry out this project. I am gracious for the invaluable advices and support I have received from colleagues such as Viktoria Lytvyn, Dr. Ratsavarinh Vongsamphanh, and Dr. Holger Lindner.

Chapter 1 Introduction

Post-translational modifications modulate enzyme activity and generate new dimensions to complex biological processes. In eukaryotes, the fate of a protein is often determined by attachment of ubiquitin moieties through covalent yet hydrolyzable enzyme mediated linkages. Polyubiquitination, the attachment of a polymeric ubiquitin chain, is classically known for its direct involvement in modulation of ubiquitin-mediated proteasomal degradation and, like the attachment of a single ubiquitin molecule, is involved in a multitude of biological functions from endocytosis to cell proliferation.

Ubiquitin is a small 76-residue protein involved in post-translational modifications that help to modulate diverse biological pathways. The ubiquitin pathway includes two biochemical phenomena: ubiquitination and deubiquitination. Ubiquitination is a reversible post-translational modification that involves the covalent linkage of ubiquitin molecules to a target protein. Deubiquitination, performed by deubiquitinating enzymes (DUBs), can cleave the isopeptide bond between ubiquitin and the site of attachment. In addition to reversing ubiquitination, DUBs are also involved in the activation of ubiquitin and ubiquitin-like modifiers through C-terminal processing of their precursors. The ubiquitin pathway is involved in a wide array of biological functions including DNA repair, signal transduction, membrane protein trafficking, endocytosis, transcription, nuclear transport, and proteolysis. The family of human enzymes ubiquitin-specific proteases (USPs) constitutes the majority of DUBs. Although biological studies over the past decade have significantly contributed to the current structural and functional inventory of DUBs, we still know little about the physiological roles and mechanisms of intermolecular interactions of most USPs.

In human, there are currently 54 USPs known with high variance in length and domain architecture. Our current understanding of the structure and function of USPs is the result of a combination of bioinformatics predictions and experimental efforts. In addition to the evident presence of a catalytic core homologous to ubiquitin C-terminal hydrolases (UCH), many USPs are found, mostly by sequence-based comparison techniques and probabilistic models such as BLAST, SMART, Pfam, and PROSITE, to harbor rather unrelated domain architectures outside the UCH domain (Nijman et al. 2005). For example, zinc-fingers are observed at the N-terminal region of some USPs (USP3, 5, 13, 16, 20, 22, 33, 44, and 49) while others have ubiquitin-associated domains (UBA) placed in tandem at the heart of the catalytic core. Despite current efforts in the annotation and characterization of these enzymes, large portions of the N- and C-terminal extensions of human USPs remain structurally and functionally un-annotated. In addition, the length of their catalytic core domains varies from approximately 300 to 800 residues due to large uncharacterized insertions which may play important physiological roles. This lacuna in functional assignment is a direct indication that even the most sensitive evolution-based sequence comparisons may not be sufficient to infer functional homologies.

The objective of this study is to develop and apply a methodology to structurally annotate previously un-annotated regions of the 54 putative human USPs. Possible functions of the resulting predictions will be inferred and discussed from predicted folds. Given the currently known and anticipated biological roles for USPs, our comprehensive structural annotation of these members of the DUB family provides important stepping stones towards elucidating their precise involvements in many human diseases.

In order to better understand the objectives and results, biological and computational background related to the subject will be presented over the next two chapters. First, the fundamentals of the prominent ubiquitin pathway will be discussed in Chapter 2. The next section, Chapter 3, confers the fundamentals of the bioinformatics aspect of our study to the reader. From sequence alignment to structure prediction, key information concerning classical and structural-bioinformatics give the reader the fundamentals necessary to comprehend our choice of method and the power of our predictions. In addition, the fundamentals of the key tool employed in our study will be presented. The published article will then be presented in Chapter 4.

Chapter 2 Biological Background

In this chapter, an overview of structural and functional properties of type 1 and type 2 ubiquitin-like (UBL) modifiers will first be presented. Next, the two fundamental processes that constitute the ubiquitin modification of proteins will be explored: ubiquitination, the attachment of ubiquitin molecules onto a protein target, and deubiquitination, their removal. Structural and functional aspects of deubiquitinating enzymes (DUBs), the focus of present study, will be presented in greater detail.

2.1 Ubiquiton: the ubiquitin superfold

Ubiquitin and UBL structures belong to the β-grasp class of structural folds. It boasts a beta-sheet wrapped around a central α-helix much like the grasp of a hand (Figure 1). The central helix flanked by two upstream and three downstream consecutive beta-sheets is characteristic to secondary structure arrangements of ubiquitin and consensus of its umbrella family: ubiquitons. A three-residue 3_{10}-helical turn often follows immediately after the central α-helix in many UBLs.

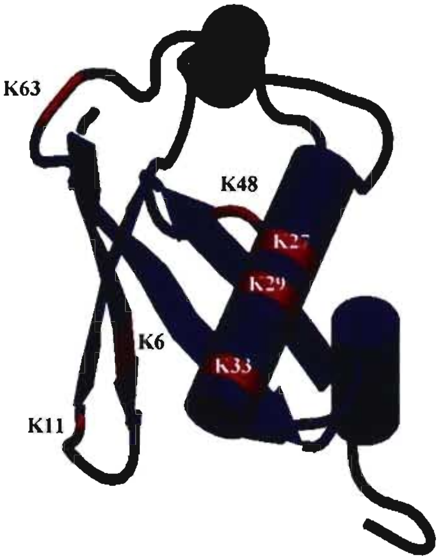

Figure 1: Cartoon representation of ubiquitin (PDB: 1AAR). All 7 lysine residues are shown (red).

In biological systems, the functional role of a protein is not only controlled by its level of expression but also by post-translational processes during which chemical modifications of the polypeptide modulate its activity, sub-cellular localization, and elimination. In eukaryotes, there exists a post-translational process widely known as ubiquitination, which is analogous to, but biologically distinct from, phosphorylation, and which consists in the addition of ubiquitin moieties to protein targets. Molecules that have a covalent mode-of-action resembling the C-terminal attachment of ubiquitin are commonly termed UBL molecules or type I ubiquitons. In contrast, there also exist UBL structures that are embedded within a larger protein, which constitute the type II class of ubiquitons.

2.1.1 Type I: ubiquitin and ubiquitin-like modifiers

Ubiquitin (Ub) is a small and highly conserved molecule abundant in eukaryotes. Its 3D structure is representative of the UBL fold, which is shared by a large number of protein sequences showing remarkable structural similarities while displaying divergent sequence homology (Kiel, Serrano 2006). This 76-residue polypeptide acts as a signaling marker that effectively guides the target protein through various biological pathways. The Ub-labeling of a target protein is usually a result of the formation of a covalent linkage between the C-terminus of Ub and the ε-amino group of a substrate lysine residue (Figure 2A). This reversible process is commonly known as ubiquitination and is governed by the sequential action of three enzymes: E1, E2 and E3, as will be described in the following section. There also exist several small proteins of close structural resemblance to ubiquitin, which have rather divergent sequence similarities and functions, but share with Ub the same 3D-structural fold and the covalent mode-of-action via a reactive C-terminus glycine (Welchman et al. 2005). The most studied type I UBL molecules are SUMO isoforms, NEDD8, ISG15 and FAT10 with functional roles ranging from transcriptional regulation (SUMO) to E3 regulation (NEDD8) to immune response (ISG15) to apoptosis (Welchman et al. 2005).

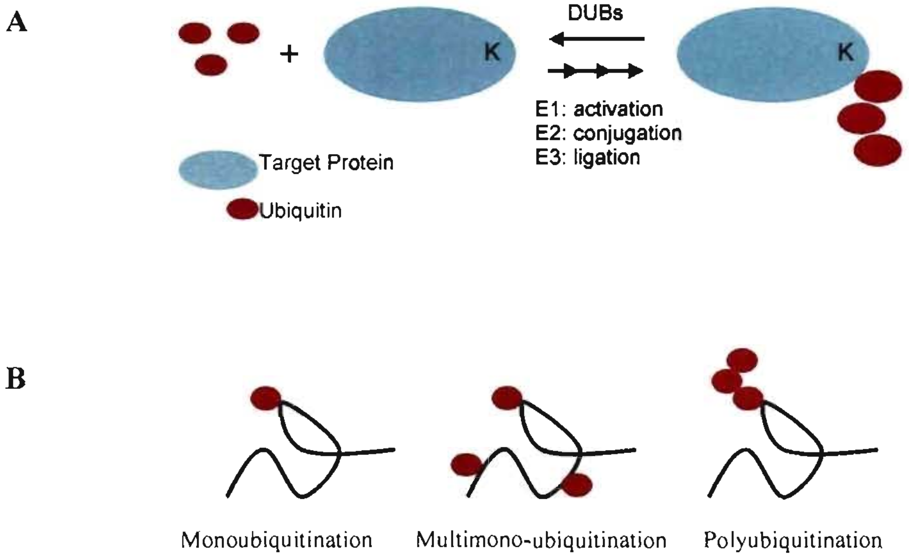

Figure 2: Target labeling by ubiquitin(s). (A) The process of ubiquitination is guided by the sequential actions of E1, E2, and E3 enzymes. (B) There exist three modes of Ub attachment: mono-, multi-, and polyubiquitination.

2.1.2 Type II: integral ubiquitin motifs

Ubiquitin-like structures also exist as integral elements of larger proteins. These genetically built-in UBL domains can be located anywhere in the sequence, although N-terminal UBL structures are prevalent in current knowledge bases. Type II UBL structures generally play integral roles in protein-protein interaction, subcellular localization, as well as intrinsic modulation of enzyme activity. Integral UBLs have also been shown to associate with the 19S subunit of the 26S proteasome primarily by interacting directly with ubiquitin-interacting motifs (UIM) of the proteasome. For example, UBL domain-containing protein Rad23, which also contains a regulatory ubiquitin-associated domain (UBA), was shown to deliver polyubiquitinated substrates to the proteasome for degradation. (Walters et al. 2004)

2.2 Ubiquitin modification pathways

The previous section described the two types of ubiquitons. We will now introduce the biochemical mechanism which underlies different mode of attachment of ubiquitin and ubiquitin-like molecules.

2.2.1 Ubiquitination

Ubiquitination is defined by the covalent linkage of Ub onto a target protein via formation of a bond typically between the C terminus (Gly76) of Ub and the ε-amino group of a substrate lysine residue. It is a three-part process mediated by sequential action of El activating, E2 conjugating, and E3 ubiquitin-protein ligating enzymes and initiates all known types of ubiquitination . The carboxyl group of Gly76 forms a thiol ester with E1, activating the C terminus of Ub for a nucleophilic attack. This activated Ub is transiently carried by an E2 to an E3, which in turn specifically recognizes and facilitates transfer of the Ub to a substrate (Pickart, Eddins 2004).

In eukaryotes, there exist three modes for ubiquitin modification, each distinguished by its characteristic pattern of attachment which triggers distinct biological pathways. First, monoubiquitination is defined by the linkage of a single Ub molecule to a single site on a target protein. Second, multi-monoubiquitination involves monoubiquitination at multiple sites. Last, polyubiquitination consists of the addition of a polyubiquitin chain constituting two or more covalently linked Ub molecules (Figure 2B).

All 7 lysine residues of the 76 amino-acid Ub offer potential linkage sites for chain extension (Figure 1). In nature, however, Lys48 and Lys63 linkages are more commonly observed. Formation of Lys48-linked polyubiquitin chains is precursor to signaling a target protein for proteasomal degradation and plays important roles in DNA repair and signal transduction. While Lys48-linked polyubiquitin chains can effectively target a protein to the proteasome for degradation, Lys63-linkage topology typically does not. The latter configuration is structurally different in comparison with Lys48-linkage and is generally recognized in DNA repair, signal transduction, and non-lysosomal degradation. (Kerscher et al. 2006) In addition, Lys63-linked chains activates NF-κB signaling pathway involved in inflammation, apoptosis, and tumorigenesis (Nijman et al. 2005).

The general response of a cell following exposure to chemical and mechanical stress is an increase in intrinsic proteolysis. Stimulation of ubiquitin-dependant proteolytic system is one of the important elements in the proteolytic cascade and is shown to respond to a host of cellular stresses caused by viral infection, heat shock, and protein damage (Wilkinson 1995). This destructive pathway is governed by proteasomal degradation of polyubiquitinated target proteins. Once covalently tagged with a Lys48-linked polyubiquitin chain, the marked protein is able to localize to the ubiquitin-interacting motif (UIM)-containing 19S subunit of the 26S proteasome and undergo proteolytic disassembly. The UIM is a short segment of about 20 residues that recognizes and binds to Ub and is often present in three consecutive copies within proteins, especially those involved in ubiquitin pathways.

Monoubiquitination and multi-monoubiquitination are commonly associated with non-degrading processes such as endocytosis, regulation of transcription, and sorting of target proteins to the vacuole (Loayza, Michaelis 1998;Soetens et al. 2001). Monoubiquitination was previously demonstrated in yeast and vertebrate cells to be sufficient to induce endocytosis of membrane-bound proteins and receptors. (Hicke, Dunn 2003;Peschard, Park 2003;Holler, Dikic 2004). Monoubiquitination is also involved in subcellular localization and in recruiting Ub binding proteins to specific interacting partners (Nijman et al. 2005).

2.2.2 Deubiquitination

The reversal of ubiquitination during which Ub moieties are excised is termed deubiquitination, a process mediated by members of the deubiquitinating enzyme (DUB) superfamily. Ubiquitin-specific proteases (USP) are cysteine proteases and form a family of ubiquitin-processing enzymes that belong to the DUB superfamily. To date, more than 54 human USPs have been reported and summarized in recent research publications and reviews (Nijman et al. 2005;Quesada et al. 2004). These enzymes possess a catalytic core homologous to that of the papain superfamily of cysteine proteases. In human, the USP catalytic core domain varies from 300 to 850 residues flanked by highly variable N- and C-terminal domain structures. Within the human USP family, homology exists primarily in the Cys-box and His-box regions harboring the catalytic residues. The highly divergent N- and C-terminal regions containing a variety of structural domains and functional motifs (e.g., zinc finger domains, EF-hand domains, UIMs) indicate an elevated structural and functional complexity within the USP family of DUBs. The large number of both validated and predicted USPs suggests that these enzymes may exhibit selectivity for specific target proteins as well as the type of Ub conjugation. The current functional understanding of USPs relies largely on sequence-based comparisons.

Despite the importance of USPs, which play key regulatory roles in a multitude of processes, our knowledge of their substrate specificity and precise mode of regulation is surprisingly scant compared to the sequence and structural landscape of this class of enzymes (Nijman et al. 2005). Human USPs harbor divergent domain structures in regions extending the catalytic core. Such diversity allows for narrow specificities for the target protein and substrate ubiquitin or ubiquitin chains as well as for modulation of activity of these enzymes (Nijman et al. 2005). The dual property of target and substrate recognition of a particular protein and ubiquitin branching, respectively, may further increase the degree of specificity of USPs. The ubiquitin pathway is well known for its key role in Ub-directed proteolysis of target proteins. Crystallographic studies have confirmed that USP14, which resides in and is regulated by the 26S proteasome, contains a previously identified N-terminus UBL domain that specifically associates with the 19S regulatory particle (Borodovsky et al. 2001;Hu et al. 2005;Nijman et al. 2005). Because of their structural dissimilarities, recognition of Lys48 and Lys63 linked poly-Ub chains often do not overlap. Amongst other branching-specific USPs, USP8 and USP14 exclusively cleave Lys48 but not Lys63 linked poly-Ub chains. Such specificity may be acquired from regions flanking the catalytic core of USPs. In fact, the ubiquitin-interacting zinc finger (Znf-UBP) is one of the factors in USP15 essential for disassembly of Ub polymers (Hetfeld et al. 2005). Interestingly, USPs are also engaged in systems that involve modifications via UBL moieties. USP21 is an example of dual recognition of both Ub and NEDD8 while USP18 is shown to cleave and maintain cellular levels of ISG15 (Gong et al. 2000;Malakhov et al. 2002).

In addition to Ub substrate specificity, regions outside the conserved catalytic core of USPs may play roles in target recognition as well. In fact, many E3 ligases are targets for USPs. NRDP1, which ubiquitinates and promotes degradation of ErbB3, an epithelial growth factor receptor (EGFR), is deubiquitinated and stabilized by USP8 (Qiu et al. 2004;Wu et al. 2004). USP8 contains a rhodanese-like domain which, in conjunction to its catalytic core, contributes to the recognition of polyubiquitinated NRDP1 (Avvakumov et al. 2006). USP7, also known as HAUSP, can stabilize the E3 Mdm2, a p53 suppressor and lead to attenuation of p53 activity (Li et al. 2004). Disruption of USP7 activity was shown to effectively suppress tumor growth (Cummins et al. 2004). Moreover, USP7 can also deubiquitinate and stabilize p53, thereby possessing dual roles in p53 regulation (Li et al. 2004). Interestingly, USP7 harbors a TRAF-like MATH domain in the N-terminal region preceding the catalytic core that recognizes and binds to both p53 and Mdm2 (Hu et al. 2006). The primary objectives of the present study are to explore the sequence-to-structure-to-function paradigm by applying structural bioinformatics tools on un-annotated regions of the 54 known human USP sequences to uncover novel information elusive to simple sequence comparison methods.

2.2.3 Deubiquitination and diseases

The implication of DUBs in human diseases has drawn significant attention and research interest to this class of enzymes in recent years. To demonstrate the extent of implications of USPs and other DUBs in human diseases, some examples are presented.

The first direct evidence of the role of ubiquitination in tumor suppressor p53 downregulation originated from human papillomavirus (HPV) studies. It was shown that degradation of tumor suppressor p53 was induced by an ubiquitin ligase complex in HPV. In fact, p53 is polyubiquitinated in cells infected by HPV leading to its proteasomal degradation (Scheffner et al. 1990;Scheffner et al. 1992). As previously noted, regulation of p53 is also modulated by Mdm2, a RING finger E3 that inhibits p53 activation function and downregulates its expression via polyubiquitination. Conversely, Mdm2 is itself regulated by p53 and these enzymes together form an auto-regulatory negative feedback loop in cell proliferation (Pickart, Eddins 2004). The USP7 enzyme, also known as herpesvirus-associated USP, or HAUSP, was originally observed to associate with herpes simplex virus-type 1 immediate-early Vmw110, a RING finger protein required for efficient initiation of viral lytic cycle (Everett et al. 1997;Sacks, Schaffer 1987). Interestingly, USP7 specifically stabilizes p53 via deubiquitination therefore inducing p53-mediated cell growth arrest and apoptosis (Cummins et al. 2004;Kim et al. 2003). In recent studies, the full length isoform of USP2 (USP2a) has been identified as an oncogenic enzyme that stabilizes fatty acid synthase (FAS) and Mdm2 via deubiquitination (Graner et al. 2004;Priolo et al. 2006;Stevenson et al. 2007). USP2a overexpression induces tumorigenesis via a mechanism opposing that of USP7. In contrast, the truncated form of USP2 promotes apoptosis when overexpressed in several cancer cell lines (Gewies, Grimm 2003). USP2 isoforms share a common catalytic core and differ only in the length of the N-terminal regions.

A significant increase in USP4 mRNA level has been reported in small cell tumors and in adenocarcinomas of the lung. USP4 has been shown to directly associate with retinoblastoma protein (pRb), a tumor suppressor protein known to be dysfunctional in a number of types of cancer, but does not exhibit deubiquitination activity (Blanchette et al. 2001). Nevertheless, A_{2A} receptor, involved in endoplasmic reticulum (ER) quality control, is a deubiquitination substrate for USP4 (Milojevic et al. 2006). In addition, reciprocity in activity between USP4 and its E3 ligase Ro52 was also identified (Wada, Kamitani 2006). Further investigation is required to understand the precise implications of USP4 in oncogenesis. USP15 also possesses a pRb-interacting motif and may be involved, in cell growth regulation via deubiquitination and thus stabilization of pRb. (Kim et al. 2003)

Emerging evidence suggests that USP8 is involved in cell proliferation by inducing degradation of EGFR and growth factor receptor tyrosine kinases (RTK) (Daviet, Colland 2007). Initially identified as a cell growth regulator that modulates the ubiquitination state of several key proteins in proliferation, USP8 is able to cleave both linear and isopeptide-linked ubiquitin chains (Naviglio et al. 1998) to rescue and recycle Ub at late endosome. Several interaction partners and substrates were identified. In mouse, Ras-guanine nucleotide exchange factor CDC25 is deubiquitinated and stabilized by USP8. An Hrs-binding protein, Hbp, which is involved in receptor endo- and exocytocis binds to USP8 via a Src homology domain 3 (SH3). USP8 was also shown to mediate T-cell anergy by preventing self-ubiquitination and degradation of the transmembrane RING finger E3 ligase GRAIL, which is closely linked to endocytic pathways. In a similar fashion, the catalytic core and rhodanese domains of USP8 were shown to bind to and stabilize NRDP1, an E3 ligase that mediates EGFR stabilization. (reviewed in Daviet, Colland 2007)

The first study demonstrating the relationship between DUBs and neurodegenerative diseases was the identification of the association of autosomal dominant point mutation in UCH-L1 with Parkinson's disease. Another example of implications of DUBs in neurological dysfunction is that in mice, a single homozygous mutation in the *ataxia* gene encoding USP14 leads to severe tremors followed by hind limb paralysis and death (D'Amato, Hicks 1965). In fact, reinstatement of USP14, for which abolishment of activity resulted in 35% decrease in Ub monomers in most of the tissue, restored Ub levels and reinstated motor functions (Anderson et al. 2005;Crimmins et al. 2006). In addition, mouse homolog of USP25 shows connection with the expression of proliferative neuroepithelial cells and post-mitotic neurons. In brain cells of Down syndrome patients, the expression of USP25 was decreased 2-fold therefore providing clues that USP25 may be involved in Down syndrome pathogenesis. (Cummins et al. 2004;Kim et al. 2003)

USPs are also involved in spermatogenesis. Genetic screening of 576 infertile and 96 fertile men revealed the link between USP9 and male infertility. A 4-bp deletion in the Y chromosome-linked USP9 (USP9Y) was determined to be responsible for the absence of sperm in the semen of azoospermic men (Sun et al. 1999).

These deubiquitination-related diseases further emphasize the current need for functional and structural annotations of USPs. The present study provides a first comprehensive examination of human USPs by looking at information beyond what the sequences alone can provide. Our computational approaches and methodology are further described in Chapter 3 and in the published paper.

Chapter 3 Structural Bioinformatics

This section will provide fundamental understanding of current tools in domain annotations and functional predictions. Sequence-based comparison methods will be briefly described. Fold recognition by threading methods will also be presented. Functional inference from a sequence-based and fold recognition methods will be discussed. The primary tool structure prediction methods used in this study, 3D-Jury, will be introduced in more detail.

3.1 Current structural and functional repositories

There exist several databases for the functions and structures of gene products. The NCBI (at the National Center for Biotechnology Information) is without question the pinnacle of all databases. Amongst other features, the GenBank database homed at NCBI contains sequence information, functional annotations, and have many useful cross-references. Sequence query in these databases is generally the first approach to annotating a new gene product with unknown functions. (Jenuth 2000) The Swiss-Prot protein knowledgebase is another well known database dedicated to proteins and contains cross-reference to a wide array of functional and structural databases including NCBI and RCSB's Protein Data Bank (PDB) (Bairoch et al. 2004). The latter database is the primary repository for atomic-resolution experimental structures of proteins and protein complexes. InterPro is the main portal to structural-functional repositories, including Pfam, SMART, ProDom, PRINTS, UniProt and ProSite. It also contains references to structural classification databases SCOP and CATH.

3.2 Structural classifications

The RCSB Protein Data Bank (PDB) is a well maintained and up-to-date collection of published structures of biomolecules determined either by NMR or X-ray crystallography. There has been an exponential growth in the number of protein structures in the PDB over the past decade. Each protein structure is classified into superfamilies of characteristic protein folds and further sorted into unique classes of closely related folds by two major classification methods: SCOP and CATH. Structural Classification of Proteins (SCOP) distributes protein domains from structural classes down to folds regardless of sequence homology (Murzin et al. 1995). SCOP is considered as a standard in classification of protein folds and relies largely on expert interpretations of protein structures. CATH is an acronym for the four main levels of hierarchical organization of structural folds: Class Architecture Topology Homology. Structural classification by CATH is a semi-automated consensus driven process guided by both sequence and structural homology information. Expert interpretations occur at the final A-stage classification via visual inspections and cross-references (Orengo et al. 1997). However, as noted in the Introduction section, the growth in the number of new folds defined by these structure classifiers has reached a stall. Therefore, it is believed that the structural classes at the present time form the basis of a great majority of protein structures in the PDB.

Ubiquitin and UBL structures belong to the β-grasp class of structural folds (SCOP entry d.15), which consists of a beta-sheet wrapped around a central α-helix much like the grasp of a hand (Figure 1). A central alpha-helix flanked by two upstream and three downstream consecutive beta-strands is characteristic to the secondary structure arrangement of ubiquitin and ubiquitons. A three-residue 3_{10}-helical turn often follows immediately after the central α-helix in many UBLs.

3.3 Prediction of function

One of the main purposes of bioinformatics is to develop and implement computational methods to annotate the function of a given, typically newly discovered, gene product. Depending on the sequence identity/homology to existing functionally annotated protein sequences, functional inference can be attempted by classical sequence-based approaches or/and via a structural bioinformatics route centered on fold detection methods.

3.3.1 From sequence to function

The classical approach to functional inference essentially relies on the pairwise alignment of the new (query) sequence with those in functionally annotated protein sequence databases (described in Section 3.1), aiming to identify hits with global or even local similarities that can provide hints about its functions. Another type of sequence-based comparison utilizes probabilistic models such as the hidden Markov chain (HMM) to compare a query sequence with a signature of probabilities for a particular sequence to occur rather than pairwise comparison of two sequences. Profile comparison is another widely used statistical approach that aligns a query sequence with a pre-determined sequence pattern characteristic to a group of sequences notably from the same family (Krogh et al. 1994). Profile building and comparison methods are more sensitive than direct pairwise alignment in detecting homology. These traditional evolutionary-based approaches to predict the function of a protein generally yield reliable results when statistically significant homology exists between the query and one or more database sequences.

The downfall of sequence-based functional assignment is that the relationship between sequence and function is neither unique nor straightforward. It is widely known that proteins exercising similar functions may exhibit divergent sequences, and vice versa. This can be explained through observations of structural similarities between two proteins even in the absence of sequence similarities, which brings about another phenomenon that similar structures may display similar functions. Therefore, structural homologies that are undetectable with sequence-based comparisons may be overlooked in classical functional annotation protocols still widely used today. Increasing cases of proteins with similar structure and functions, but undetectable sequence similarity have driven development of more sensitive methods for structure-function prediction (Brenner et al. 1998;Martin et al. 1998).

3.3.2 From sequence to structure to function

The fold recognition concept, initially implemented to identify analogous proteins with undetectable sequence homologies, has gained popularity in the 1990s. Fold detection appeared as a necessary tool due to the observation of structurally related proteins with dissimilar primary sequences. This may have arisen from convergent evolution of two initially unrelated genes due to pressure exercised by external factors from habitats of two different species, or from rising demand of a cell to perform a given task. Fold detection in cases of low sequence homology thus opened new avenues for function inferences, however with the caveat that similarly folded protein structures with completely different biochemical functions do exist.

The traditional approach to fold prediction is based on sequence-sequence or sequence-profile comparisons. Homology modeling for example utilizes the evolutionary information obtained from sequence alignments and an associated scoring matrix to predict the spatial arrangement of each residue. An immediate extension to the sequence-sequence comparison approach has emerged with the implementation of threading methods. This technique utilizes a template structure to compute position-specific alignment scores based on iterative calculation of substitution scores by replacing the side-chain of a residue with all other 19 naturally occurring side-chains. Another approach to fold prediction is based solely on a first-principle physical, but time-consuming, treatment. These *ab initio* methods take advantage of physical properties of atoms in order to calculate the free energy over an ensemble of protein conformations and to simulate protein folding. Current limitations in computation power, however, do not allow sufficient sampling flexibility to effectively predict protein folds via *ab initio* methods. A final class of protein structure prediction methods, termed meta-predictors, applies statistical methods to improve the accuracy of a collection of protein structure predictions over that of individual methods.

Meta-prediction is a simple yet powerful approach (Ginalski et al. 2005) that utilizes a diverse collection of prediction algorithms as the basis for arriving at a consensus 3D-structural prediction for a query protein sequence. Meta Servers belong to an online framework of meta-predictors that gather or implement, and then analyze models predicted by individual servers or methods. The principle behind meta-predictors relies on observations that the most abundant low-energy conformation (from simulated structures generated by *ab initio* prediction protocols) is closer to the native structure than the conformation with the lowest energy (Ginalski et al. 2003). This statement translates into the philosophy behind meta-predictors by inferring that most abundant high-scoring models are closer to the native structure than the model with highest score (Ginalski et al. 2003). The consensus approach has experienced great success in this domain and is included in a series of biennial benchmark studies, CASP, a large-scale experiment launched in 1994 to assess protein structure prediction methods and presently at version 7. All results obtained by the published CASP6 (version 6 of CASP) experiment, which was completed for the period ending in December 2004, indicate that meta-predictors are more accurate than any independent fold recognition methods (Wang et al. 2005).

There exists several fully automated Meta Servers for consensus prediction. Historically, the first server, Pcons, implemented a neural network to uniformly scale the confidence score of models from various methods based on the expected accuracy of individual models (Ginalski et al. 2005;Lundstrom et al. 2001). This approach outperformed any individual method included by Pcons by generating ~8-10% more correct predictions and with a significantly higher specificity. A second consensus prediction is 3D-Jury which, unlike Pcons, solely relies on the statistical significance of predicted models (Ginalski et al. 2003).

3.4 3D-Jury - a consensus fold recognition server

3D-Jury (Ginalski et al. 2003) is a simple yet powerful meta-predictor and is currently part of continuously-run structure prediction benchmark LiveBench and in CAFASP evaluation of fully-automated fold prediction servers (Bourne 2003;Rychlewski, Fischer 2005). By including results from other meta-predictors in its consensus calculations, 3D-Jury has also earned the name “meta-meta-predictor”. During the course of this project, the 3D-Jury server included the Pcons meta-server for fold prediction (Wallner et al. 2003). In a similar fashion as in clustering of similar structures from *ab initio* simulations, 3D-Jury identifies, through a simple normalized summation over similarity scores between each model and all other input models, the best structure at the center of all predicted models. It thus can be considered as a non-energetic prediction method, since model ranking relies on the repeated occurrence of low-energy models rather than on the scores of these models from independent prediction methods. A similarity score between a pair of 3D-models is defined as the number of corresponding Cα-atom pairs within 3.5 Å of each other after optimum 3D-superimposition of the model. A similarity score of 40 represents a threshold for reliable structural predictions, based on the observation that it corresponds to a ~90% confidence that the underlying 3D-structures belong to the same class of protein fold (Ginalski et al. 2003). Two modes are implemented in 3D-Jury: 1) The all-model mode considers all predicted models, while 2) the best-model mode discards all but the model from each server with the best similarity score with all other models. The set of models generated from the selected mode are then used for calculating the 3D-Jury score. At the time of this study, results from 15 servers from sequence-based comparisons to threading methods to meta-servers were gathered in the 3D-Jury system (http://bioinfo.pl/meta/). Data collected from servers harboring threading algorithms (mGenThreader, INUB, Sam-T02, FUGUEv2, and 3D-PSSM), highly accurate sequence-only analyses (FFAS(03), Meta-BASIC(3), BasicDist, ORFeus2, Psi-BLAST, and Superfamily), and secondary structure predictors (PROFsec, PSI-PRED) were used as jury.

In LiveBench6 experiments, 3D-Jury demonstrated high sensitivity on difficult targets, which are outliers from structural alignment of PDB entries within the same fold class (Rychlewski et al. 2003). In the same study, 3D-Jury was shown to produce the highest number of correct predictions in both difficult and easy targets using variations in the subset of model predictions.

Chapter 4 Published Article

WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS

PREDICTION REPORT

High incidence of ubiquitin-like domains in human ubiquitin-specific proteases

Xiao Zhu,[1,2] Robert Ménard,[1,2] and Traian Sulea[1]*

[1] Biotechnology Research Institute, National Research Council of Canada, Montreal, Quebec H4P 2R2, Canada

[2] Department of Biochemistry, Université de Montréal, Montreal, Quebec H3C 3J7, Canada

ABSTRACT

Ubiquitin-specific proteases (USPs) emerge as key regulators of numerous cellular processes and account for the bulk of human deubiquitinating enzymes (DUBs). Their modular structure, mostly annotated by sequence homology, is believed to determine substrate recognition and subcellular localization. Currently, a large proportion of known human USP sequences are not annotated either structurally or functionally, including regions both within and flanking their catalytic cores. To extend the current understanding of human USPs, we applied consensus fold recognition to the unannotated content of the human USP family. The most interesting discovery was the marked presence of reliably predicted ubiquitin-like (UBL) domains in this family of enzymes. The UBL domain thus appears to be the most frequently occurring domain in the human USP family, after the characteristic catalytic domain. The presence of multiple UBL domains per USP protein, as well as of UBL domains embedded in the USP catalytic core, add to the structural complexity currently recognized for many DUBs. Possible functional roles of the newly uncovered UBL domains of human USPs, including proteasome binding, and substrate and protein target specificities, are discussed.

Proteins 2007; 69:1–7.

Key words: consensus fold recognition; deubiquitination; proteasome; ubiquitin; UBL; USP.

INTRODUCTION

Post-translational ubiquitination of proteins in eukaryotes governs cellular activities ranging from selective protein degradation by proteasomes to membrane protein trafficking, signal transduction, transcription, nuclear transport, autophagy, and immune responses.[1–4] Protein ubiquitination is catalyzed by the sequential action of E1, E2, and E3 enzymes that activate and transfer ubiquitin or ubiquitin-like modifiers to the ϵ-amino group of an internal lysine residue of target proteins.[5,6] Ubiquitination is a reversible process. The isopeptide bond between ubiquitin and a substrate protein, or between ubiquitin molecules in a polyubiquitin chain, can be cleaved by deubiquitinating enzymes (DUBs), which are also responsible for the activation of ubiquitin and ubiquitin-like modifiers by C-terminal processing of their precursors.[7] A large number of DUBs have been discovered and represent an emerging class of ubiquitin pathway regulators, predominantly from eukaryotes,[8] but also of bacterial and viral origins.[9–11]

New insights into molecular structures, biochemical activities, substrate specificities and functions have been gained for the current inventory of DUBs over the past decade.[7,8] Most known cellular DUBs are cysteine proteases, including those from the ubiquitin-specific protease (USP) structural class, which represents the bulk (over 50) of DUBs encoded in the human genome.[8,12] Little is known about the physiological function of most human USPs, and specific substrates remain elusive. The current view is that the modular, multidomain architecture of USPs contributes to their specificity with respect to the type of ubiquitin polymer and modifier, but perhaps more importantly, to the target protein part of the substrate.[8] Human USPs have highly variable amino acid sequences upstream and/or downstream of the catalytic core. A number of domains have been

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

*Correspondence to: Dr. Traian Sulea, Biotechnology Research Institute, National Research Council of Canada, 6100 Royalmount Avenue, Montreal, Quebec H4P 2R2, Canada. E-mail: [information retirée / information withdrawn]

Received 9 December 2006; Revised 19 February 2007; Accepted 19 March 2007

Published online 27 June 2007 in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/prot.21546

annotated in these regions based on sequence homology,[8,12] some already confirmed experimentally, for example, the TRAF-like domain of human USP7, the DUSP domain of human USP15, and the CS domain of human USP19. However, a large proportion of the N- and C-terminal extensions of human USPs remain structurally and functionally unannotated. Also, the size of their catalytic core domains varies from ~300 to 800 residues due to large sequences uncharacterized structurally, which may play functional roles.

Given the currently known and expected important cellular roles of USPs, a detailed structural annotation of individual family members of this class of DUBs is an important step toward elucidating their molecular functions in human health and disease. On this account, we have subjected the currently unannotated content of human USP family to advanced structural bioinformatics techniques. The most impressive finding of this prediction exercise is the abundance of ubiquitin-like (UBL) domains in this family of enzymes, both within and outside USP catalytic core domains. The newly uncovered UBL domains are likely to play important functional roles toward the substrate and target protein specificities of human USPs.

MATERIALS AND METHODS

Sequences of the currently known human USPs corresponding to the C19 family of the MEROPS peptidase database (http://merops.sanger.ac.uk/) were collected from the GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) and SwissProt (http://www.expasy.org/sprot/) databases. Only one sequence was selected from those of multiple isoforms reported for some USPs (generally nearly identical mutation isoforms, otherwise the longest sequence was selected), thus leading to a nonredundant set of 54 distinct human USP sequences (see Supplementary Material). The boundaries of their catalytic core, as well as all their currently annotated domains outside this core domain were obtained from the Pfam (http://www.sanger.ac.uk/Software/Pfam/) and InterPro (http://www.ebi.ac.uk/interpro/) databases and confirmed, whenever available, with actual structures retrieved from the Protein Data Bank (PDB, http://www.rcsb.org/). The remaining unannotated sequence content was therefore defined by the sequences flanking or between the currently annotated domains, as well as inserted in the catalytic core. In the latter case, locating such insertions required (i) a multiple sequence alignment of all 54 catalytic core sequences, which was performed with the MAFFT5 algorithm,[13] and (ii) comparisons with the minimal catalytic core domain delineated by its available crystal structures from several human USPs (with PDB IDs): 2 (2HD5), 7 (1NB8, 1NBF, 2F1Z), 8 (2GFO), and 14 (2AYN, 2AYO).

Structural domain detection of the currently unannotated content of the human USP family was carried out at the Structure Prediction Meta Server (http://meta.bioinfo.pl/), which assembles state-of-the-art fold recognition methods, and provides a consensus scoring of the three-dimensional structure predictions generated for a given query sequence by independent algorithms, using the 3D-Jury meta-predictor.[14] Short sequence stretches (<40 residues) were not considered. Overly long contiguous sequences (>800 residues) were split into shorter fragments prior to fold recognition calculations. This splitting was done in two ways: (i) generating three equal-length sequences corresponding to the N- and C-terminal halves plus the central region of the same length, and (ii) following the consensus predictions of domain boundaries generated by the Meta-DP metaserver (http://meta-dp.cse.buffalo.edu/).[15] Considering the possibility of embedded domain folds, newly identified domains were excised out of the original query sequence (typically longer), and the resulting flanking regions were merged and subjected to a new round of fold detection. Finally, the excised sequences of all newly mapped domains were resubmitted to the Structure Prediction Meta Server to obtain the final template ranking, reliability indicators, query-to-template sequence alignments and secondary structure predictions.

The reliability of fold assignment was based primarily on the 3D-Jury confidence score, which was calculated using the standard settings under which the score was found to correlate to the number of correctly predicted residues.[16] Accordingly, a confidence threshold of 50 for the 3D-Jury score translates into a prediction reliability of over 90%. For shorter sequences (<100 residues), the 3D-Jury confidence cut-off was lowered to 40. A qualitative evaluation of the query-to-template sequence and secondary structure alignments was also carried out to support the assessment of each top-ranked structural assignment.

Secondary structure predictions were based on four methods: PROFsec,[17] PSI-PRED,[18] and SAM-T02 with DSSP and STRIDE alphabets.[19] A consensus was then derived for each sequence by (i) majority voting over all four methods for α-helices and β-strands, and (ii) SAM-T02 predictions of G-helices, a secondary structure not available from the other prediction methods. The multiple sequence alignment of the identified UBL domains was assembled starting from individual query-to-template sequence alignments top-ranked by 3D-Jury consensus fold recognition. This preliminary alignment was further refined by: (i) considering the structure-based sequence alignment between the top-ranked UBL templates, which was generated with the Expresso (3D-Coffee) program,[20] and (ii) minor local improvements in the sequence and secondary structure alignments among predicted UBL domains. Sequence homology-based clustering of predicted UBL domains of USPs was derived with the Clustal

W program[21] using the PAM350 scoring matrix, given the sequence divergence of the UBL fold.

RESULTS

One approach toward extending the current sequence-homology-based domain annotation of human USPs is to detect structural relationships that have only remote or no underlying sequence homology. This is *the* objective of fold recognition methods. Thus, we subjected the unannotated sequence content of the human USP family to the consensus protein structure prediction method 3D-Jury.[14,16] This widely used meta-predictor performs consensus scoring over the 3D models generated by state-of-the-art fold recognition algorithms, and ranked as a top-performer at the latest CASP, CAFASP, and LiveBench prediction contests.[22,23] We also have recently employed 3D-Jury to predict the USP-like structure and infer the deubiquitinating activity for the SARS coronavirus papain-like protease,[24,25] predictions which were experimentally confirmed both functionally and structurally.[26–28]

One of the most interesting results stemming out of this analysis was the prediction of ubiquitin-like (UBL) domains in an unexpectedly high number of human USPs (Fig. 1A). These UBL domains were predicted with high reliability as judged by the statistically significant 3D-Jury scores obtained for the corresponding USP sequences against numerous UBL templates (see Supplementary Material). Consistent with the fold recognition data, the newly identified UBL domains follow the consensus secondary structure and the common fingerprint sequence characteristic to the ubiquitin superfold (Fig. 1B).[29]

As shown schematically in Figure 1A, the previously unannotated UBL domains detected for various human USPs by our structural bioinformatics analysis are present both inside and outside their catalytic core domains. Ubiquitin-like domains nested inside catalytic core domains are found in the human USPs 4, 6, 11, 15, 19, 31, 32, and 43. In all these enzymes, the UBL domain insertion occurs at highly homologous positions, specifically, in the middle of the circularly permuted Zn-finger-like domain, itself nested within the catalytic core between the two sub-domains of the papain-like fold.[30] The nested UBL domain would be inserted in these USPs between the β-strand and the α-helix that are grafted onto the four-stranded β-ribbon of the circularly permuted Zn-finger and are utilized for its attachment to the C-terminal sub-domain of the papain-like fold (Fig. 1C), as observed in the crystal structures of several USPs.[31–33] In each case, the inserted UBL domain is directly followed by a region of about 170-240 residues (depending on the enzyme) before the remainder of the circularly permuted Zn-finger-like fold (Fig. 1A). No fold similarity could be detected for any of these regions, which, for most parts, lack predicted secondary structure elements. An interesting variation is observed for USP19, where an annotated MYND Zn-finger domain of about 45 residues is intercalated immediately after the nested UBL domain and before the large, mostly unstructured, region.

The remainder of the newly identified UBL domains are located outside the boundaries of the catalytic core domain (Fig. 1A). Ubiquitin-like domains N-terminal to the catalytic core are detected in human USPs 4, 9X, 9Y, 11, 15, 24, 32, 34, and 47. Thus, USPs 4, 11, 15, and 32 feature two UBL domains, one inside and the other one immediately upstream to the catalytic core domain. Interestingly, the predicted N-terminal UBL domain in all these four USPs is preceded by a DUSP domain.[34] The single N-terminal UBL domains of USPs 9X, 9Y, 24, and 34 are predicted to be flanked on both sides by all-α-helical domains (not shown). Multiple UBL domains were detected in the C-terminal extensions relative to the catalytic core of human USPs 7 (four domains), 40 (two domains), and 47 (three domains; a fourth UBL domain is predicted upstream to the catalytic core).

DISCUSSION

Based on primary sequence homology, UBL domains have been previously detected only in the N-terminal part of human USP14,[35] and in the C-terminal end of human USP48.[12] In the former case, the solution NMR structure of the UBL domain of mouse USP14 (PDB ID: 1WGG; 97% sequence identity to the human domain) confirms this structural assignment. The exquisite promiscuity of the ubiquitin superfold to variations in primary sequence,[29] may have precluded the detection of most UBL domains by simple applications of standard homology tools such as PSI-BLAST,[36] possibly leading to their under-representation in the currently available public annotations of USPs. Supporting this idea, the only other previously reported UBL domain of a human USP, that of USP9Y, resulted from a fold recognition-based annotation study targeted to the male-specific region of the human Y chromosome.[37]

The present structural bioinformatics analysis of the currently unannotated content of the entire human USP family significantly augments the existing annotation with the addition of 26 UBL domains from 15 distinct human USPs. Thus, the UBL domain can be regarded as the most frequently occurring domain in the human USP family, after the characteristic protease core domain.[8,12] We cannot exclude the possibility that a few other UBL domains, perhaps more remotely related to the currently known members of the ubiquitin superfold,[29] have escaped our fold detection employing the *existing* best-performing algorithms and the *current* PDB content. For

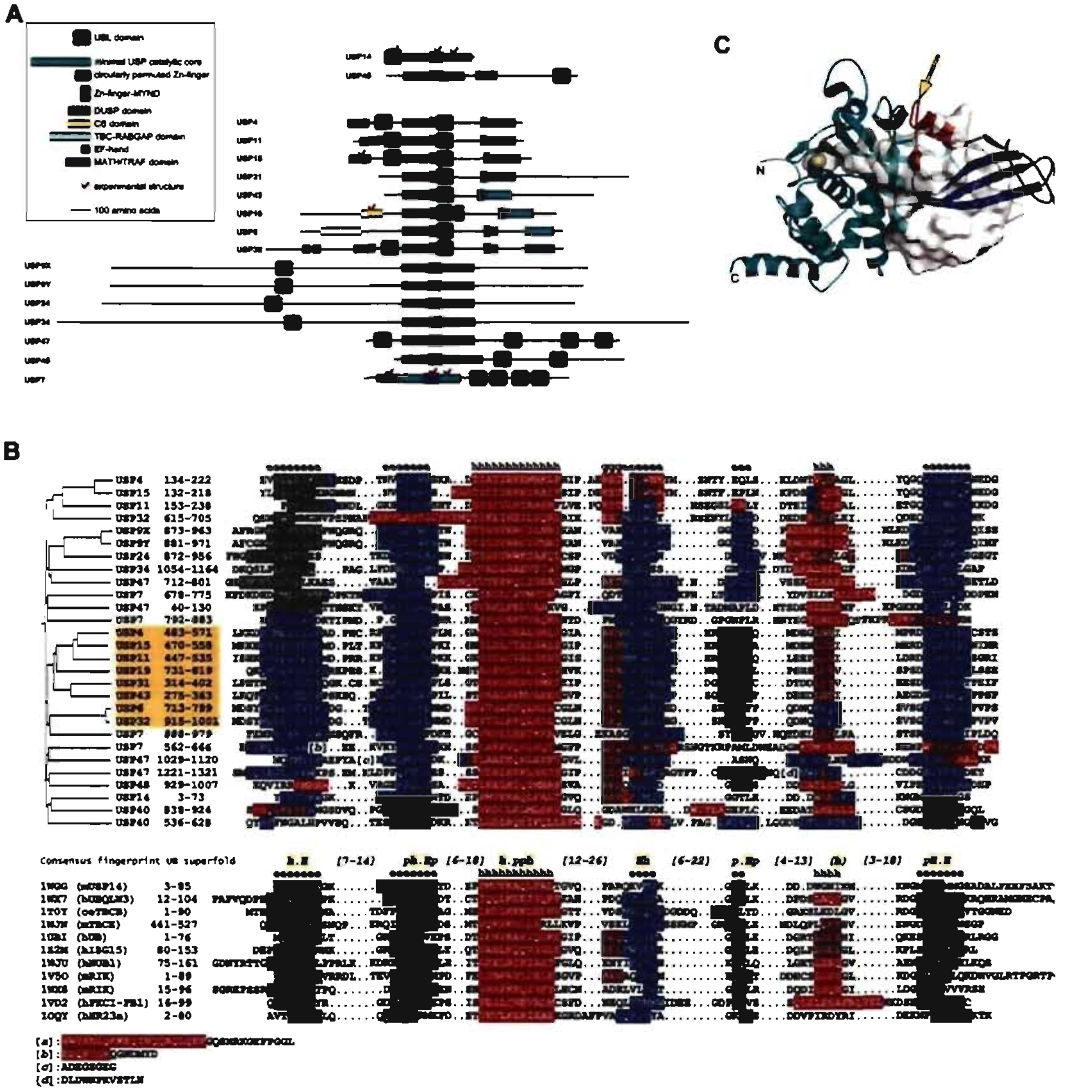

Figure 1

Novel UBL domains of human USPs predicted by structural bioinformatics. (A) Schematic domain organization of human USPs with predicted UBL domains (approximate scale). Note the split in the circularly permuted Zn-finger-like domain (blue) and the catalytic core (cyan) due to sizable insertions in some family members. Other domains of these USPs, according to their currently available public annotations, are also shown (see insert). Human USPs 14 and 48, with UBL domains previously annotated based on sequence homology, are shown at the top. (B) Sequence and secondary structure alignment between UBL domains predicted for human USPs (upper part) and selected members from the ubiquitin superfold (lower part, identified primarily by their PDB codes). See Materials and Methods for details, and Supplementary Material for other UBL structures identified by consensus fold recognition as statistically significant templates for the newly detected UBL domains, including the full query-to-template assignment. Secondary structure elements, predicted for USP sequences, and observed for the other UBL domains, are highlighted in blue – β-strand, red – α-helix, brown – G-helix, and violet – ambivalent α/β predictions. Consensus secondary structures over the query and template alignments are indicated by e – β-strand, h – α-helix, and g – G-helix. The common fingerprint sequence of the ubiquitin superfold is taken from Ref. 29, where conserved hydrophobic residues are denoted by h (90% conserved) and H (100% conserved) and conserved polar residues are denoted by p (90% conserved). A sequence homology-based clustering of the UBL domains from human USPs is shown on the left. The UBL domains inserted in USP catalytic core domains are highlighted in yellow. (C) Location of UBL domain insertion (arrow) in the minimal USP catalytic core domain, exemplified here from human USP7 (PDB code 1NBF). The papain-like protease fold is colored in cyan. The nested circularly permuted Zn-finger-like domain is in blue, except for its β3-α1 region (in red) grafted onto the β-ribbon and attached to the papain-like domain. The catalytic Cys side chain is rendered as space-filled model. Bound ubiquitin aldehyde is displayed with translucent molecular surface.

example, several other putative UBL domains from USPs 7, 40, and 47 were detected with scores below the significance threshold and therefore not reported here. The occurrence of multiple UBL domains per USP protein, as well as of UBL domains embedded in the USP catalytic core (hence comprising three distinct folds in a nested rather than successive type of assembly) add to the structural complexity that is currently being recognized for many DUBs.

The marked presence of UBL domains in human USPs is likely to have profound functional consequences in this class of enzymes, primarily via (i) modulation of enzymatic activity and specificity, and (ii) recruitment of nonsubstrate partners. In principle, due to structural similarities to ubiquitin, modulation of USP enzymatic activity by UBL domains could arise from the trivial action of competing directly with ubiquitinated substrates for the ubiquitin binding site on the Zn-finger-like domain, thus acting as auto-inhibitory domains. The newly identified UBL domains may also affect the enzymatic specificities of their USPs toward the substrate protein part or toward the ubiquitin part of the substrate conjugate. Owing to their proximity to the Zn-finger-like domain implicated directly in ubiquitin docking,[31,33] the UBL domains nested within the catalytic core (Fig. 1A and 1C) may affect the specificities of these USPs with respect to the degree (mono/di/tetra/poly) and type (e.g., K48/K63, branched/linear) of polyubiquitination, or to the type of modifier (e.g., ubiquitin, ISG15, NEDD8). For instance, an intact metal-chelating Zn-finger-like domain of human USP15, which features a nested UBL domain, is essential for degradation of K48-branched polyubiquitin chains but not for hydrolysis of the ubiquitin-GFP fusion.[38] This, in light of the fact that USP7 can degrade polyubiquitin substrates,[39] in the absence of the metal-chelating ability of its Zn-finger-like domain *and* of a nested UBL domain, suggests that the metal center might stabilize or position the nested UBL domain in order to allow polyubiquitin recognition by some USPs.

More broadly, UBL domains can engage in specific interactions with domains of both substrate and nonsubstrate protein targets. In the latter case, such specific interactions will determine noncatalytic properties of USPs such as localization, trafficking, and participation in intracellular and signaling pathways, although they can also affect the USP enzymatic activity and specificity. For example, human USP14 and its yeast homolog Ubp6 bind through their N-terminal UBL domain to the 26S proteasome, which also results in a dramatic increase of their catalytic activity.[32,40–42] Proteasome-associated DUBs can act catalytically to remove (poly)ubiquitin before proteasomal degradation, thus serving in the editing of poorly targeted substrates and ubiquitin recycling, or noncatalytically to delay proteasomal degradation and regulate both the nature and magnitude of proteasomal activity.[43] It is tempting to speculate that other human USPs would also have the ability to bind directly to proteasome subunits via some of their UBL domains identified here. The UBL domain of mouse USP14, as well as other UBL domains known to associate with proteasome, for example, from human ubiquilins 2 and 3, or DNA damage repair protein Rad23A,[44–46] were retrieved among the high scoring structural templates for newly mapped UBL domains of several USPs (see Supplementary Material). In this context, it is also worth to note the association of human USP15 with the COP9-signalosome (CSN), which has subunits similar to components of the 26S proteasome lid complex.[38] Detection of the UBL fold in USP15 suggests that the mechanism of CSN-association of USP15 might resemble proteasome-association of USP14, that is, direct interactions via UBL domains.

Functions other than proteasome and CSN binding can also be anticipated, at least in some cases, given the notorious functional variability within the ubiquitin domain superfold, comprising ubiquitin, ubiquitin-like modifiers and internal UBL structures.[29] Generally, the interaction surface seems not to be conserved within the ubiquitin superfold and almost every element of the fold is used in protein recognition, although interactions with one protein family tend to use the same surface.[29,47–49] A vast repertoire of ubiquitin-binding domains (UBDs) is known to interact with members of the ubiquitin superfold. Ubiquitin-dependent signaling pathways include UIM, UBA, UBL, CUE, GAT, GLUE, and various types of Zn-fingers such as ZnF-A20, NZF, and ZnF-UBP, among other UBDs.[48] Outside the ubiquitin pathways, small GTPases represent the prevalent fold interacting with UBL domains. The presence of UBDs can be diagnostic of putative substrate or nonsubstrate protein targets, largely unknown for most UBL domain-containing human USPs.

While the ubiquitin-like fold was abundantly detected in the relatively large family (54 members) of human USPs, its presence in USPs from older eukaryotes may indicate that UBL domains represent important functional features that were conserved during evolution in this class of enzymes. Indeed, a fold recognition analysis of the unannotated sequence content of the USP family from the *Saccharomyces cerevisiae* yeast (16 enzymes, Ubp1 to Ubp16), detects UBL domains in Ubp12 and Ubp15, the yeast homologs of human USP15 and USP7, respectively, in addition to the known UBL domain in Ubp6 (see Supplementary Material). Another compelling example suggesting the importance of the UBL domain as a key structural and putatively regulatory module in the USP class of DUBs is provided by the SARS coronavirus papain-like protease. This viral enzyme, whose primary function is in viral replication via polyprotein processing, not only acquired the USP molecular architecture and deubiquitinating activity common to the corresponding host cell enzymes,[11,26,27] but it also features a UBL domain (albeit about 20 residues smaller than ubiquitin).[28]

The UBL domains detected in human USPs do not show, in the sequence following their C-terminal β-strand, similarity to the ubiquitin precursor processing site, which occurs after the ubiquitin C-terminal sequence ${}_{73}$LRGG${}_{76}$. We therefore conclude that they are internal noncleavable UBL domains. However, a structural relationship to ubiquitin has been recently proposed for human USP1, which is inactivated by auto-cleavage after an internal ubiquitin-like di-glycine motif.[50] The cleavage occurs in the middle of a 140-residue insertion in the minimal USP catalytic core domain, based on the available crystal structures of USPs 2, 7, 8, and 14. However, we could not detect the UBL fold in the 70-residue inserted sequence preceding the processing site. It is possible that in the case of USP1, key substrate interactions between its catalytic groove S_6-S_1 subsites and its residues ${}_{666}$IGLLGG${}_{671}$ preceding the cleavage site, which are homologous to the ubiquitin C-terminal sequence ${}_{71}$LRLRGG${}_{76}$, are sufficient for cleavage without the need of additional interactions from the rest of the ubiquitin fold. Like the acquisition of noncleavable UBL domains, the utilization of a cleavable internal ubiquitin-like C-terminal motif in USP1 provides another fascinating example for the reuse of structural elements specific to the ubiquitin signaling pathways towards increasing their own regulatory capabilities and functional diversity. The structural bioinformatics analysis reported here provides valuable information that can spur further structure-function characterization studies in this class of deubiqutinating enzymes.

ACKNOWLEDGMENTS

We thank Dr. Holger A. Lindner for useful discussions and critical reading of the manuscript. X.Z. thanks Dr. Enrico O. Purisima for securing financial support to carry out this work.

REFERENCES

1. Glickman MH, Ciechanover A. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. Physiol Rev 2002;82:373–428.
2. d'Azzo A, Bongiovanni A, Nastasi T. E3 ubiquitin ligases as regulators of membrane protein trafficking and degradation. Traffic 2005;6:429–441.
3. Haglund K, Dikic I. Ubiquitylation and cell signaling. EMBO J 2006;24:3353–3359.
4. Welchman RL, Gordon C, Mayer RJ. Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol 2005;6:599–609.
5. Hershko A, Ciechanover A. The ubiquitin system. Annu Rev Biochem 1998;67:425–479.
6. Pickart CM, Eddins MJ. Ubiquitin: structures, functions, mechanisms. Biochim Biophys Acta 2004;1695:55–72.
7. Amerik AY, Hochstrasser M. Mechanism and function of deubiquitinating enzymes. Biochim Biophys Acta 2004;1695:189–207.
8. Nijman SMB, Luna-Vargas MPA, Velds A, Brummelkamp TR, Dirac AMG, Sixma TK, Bernards R. A genomic and functional inventory of deubiquitinating enzymes. Cell 2005;123:773–786.
9. Zhou H, Monack DM, Kayagaki N, Wertz I, Yin J, Wolf B, Dixit VM. Yersinia virulence factor YopJ acts as a deubiquitinase to inhibit NF-kappaB activation. J Exp Med 2005;202:1327–1332.
10. Misaghi S, Balsara ZR, Catic A, Spooner E, Ploegh HL, Starnbach MN. Chlamydia trachomatis-derived deubiquitinating enzymes in mammalian cells during infection. Mol Microbiol 2006;61:142–150.
11. Sulea T, Lindner HA, Menard R. Structural aspects of recently discovered viral deubiquitinating activities. Biol Chem 2006;387:853–862.
12. Quesada V, Diaz-Perales A, Gutierrez-Fernandez A, Garabaya C, Cal S, Lopez-Otin C. Cloning and enzymatic analysis of 22 novel human ubiquitin-specific proteases. Biochem Biophys Res Commun 2004;314:54–62.
13. Katoh K, Kuma Ki, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 2005;33:511–518.
14. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
15. Saini HK, Fischer D. Meta-DP: domain prediction meta-server. Bioinformatics 2005;21:2917–2920.
16. Ginalski K, Rychlewski L. Detection of reliable and unexpected protein fold predictions using 3D-Jury. Nucleic Acids Res 2003; 31:3291–3292.
17. Rost B. Protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218.
18. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
19. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 2006;53:491–496.
20. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 2006;34:W604–W608.
21. 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;31:3497–3500.
22. Ginalski K. Comparative modeling for protein structure prediction. Curr Opin Struct Biol 2006;16:172–177.
23. Fischer D. Servers for protein structure prediction. Curr Opin Struct Biol 2006;16:178–182.
24. Sulea T, Lindner HA, Purisima EO, Menard R. Deubiquitination, a new function of the severe acute respiratory syndrome coronavirus papain-like protease? J Virol 2005;79:4550–4551.
25. Sulea T, Lindner HA, Purisima EO, Menard R. Binding site based classification of coronaviral papain-like proteases. Proteins 2006; 62:760–775.
26. Barretto N, Jukneliene D, Ratia K, Chen Z, Mesecar AD, Baker SC. The papain-like protease of severe acute respiratory syndrome coronavirus has deubiquitinating activity. J Virol 2005;79:15189–15198.
27. Lindner HA, Fotouhi-Ardakani N, Lytvyn V, Lachance P, Sulea T, Menard R. The papain-like protease from the severe acute respiratory syndrome coronavirus is a deubiquitinating enzyme. J Virol 2005;79:15199–15208.
28. Ratia K, Saikatendu KS, Santarsiero BD, Barretto N, Baker SC, Stevens RC, Mesecar AD. Severe acute respiratory syndrome coronavirus papain-like protease: structure of a viral deubiquitinating enzyme. Proc Natl Acad Sci USA 2006;103:5717–5722.
29. Kiel C, Serrano L. The ubiquitin domain superfold: structure-based sequence alignments and characterization of binding epitopes. J Mol Biol 2006;355:821–844.

30. Krishna SS, Grishin NV. The finger domain of the human deubiquitinating enzyme HAUSP is a zinc ribbon. Cell Cycle 2004; 3:1046–1049.
31. 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;111:1041–1054.
32. Hu M, Li P, Sing L, Jeffrey PD, Chenova TA, Wilkinson KD, Cohen RE, Shi Y. Structure and mechanisms of the proteasome-associated deubiquitinating enzyme USP14. EMBO J 2005;24:3747–3756.
33. Renatus M, Parrado SG, D'Arcy A, Eidhoff U, Gerhartz B, Hassiepen U, Pierrat B, Riedl R, Vinzenz D, Worpenberg S, Kroemer M. Structural basis of ubiquitin recognition by the deubiquitinating protease USP2. Structure 2006;14:1293–1302.
34. de Jong RN, AB E, Diercks T, Truffault V, Daniels M, Kaptein R, Folkers GE. Solution structure of the human ubiquitin-specific protease 15 DUSP domain. J Biol Chem 2006;281:5026–5031.
35. Wyndham AM, Baker RT, Chelvanayagam G. The Ubp6 family of deubiquitinating enzymes contains a ubiquitin-like domain: SUb. Protein Sci 1999;8:1268–1275.
36. 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;25:3389–3402.
37. Ginalski K, Rychlewski L, Baker D, Grishin NV. Protein structure prediction for the male-specific region of the human Y chromosome. Proc Natl Acad Sci USA 2004;101:2305–2310.
38. Hetfeld BKJ, Helfrich A, Kapelari B, Scheel H, Hofmann K, Guterman A, Glickman M, Schade R, Kloetzel PM, Dubiel W. The zinc finger of the CSN-associated deubiquitinating enzyme USP15 is essential to rescue the E3 ligase Rbx1. Curr Biol 2005; 15:1217–1221.
39. Canning M, Boutell C, Parkinson J, Everett RD. A RING finger ubiquitin ligase is protected from autocatalyzed ubiquitination and degradation by binding to ubiquitin-specific protease USP7. J Biol Chem 2004;279:38160–38168.
40. 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;20:5187–5196.
41. 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;10:495–507.
42. Borodovsky A, Ovaa H, Kolli N, Gan-Erdene T, Wilkinson KD, Ploegh HL, Kessler BM. Chemistry-based functional proteomics reveals novel members of the deubiquitinating enzyme family. Chem Biol 2002;9:1149–1159.
43. Hanna J, Hathaway NA, Tone Y, Crosas B, Elsasser S, Kirkpatrick DDS, Leggett DS, Gygi SP, King RW, Finley D. Deubiquitinating enzyme Ubp6 functions noncatalytically to delay proteasomal degradation. Cell 2006;127:99–111.
44. 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;41:1767–1777.
45. Ko HS, Uehara T, Tsuruma K, Nomura Y. Ubiquilin interacts with ubiquitylated proteins and proteasome through its ubiquitin-associated and ubiquitin-like domains. FEBS Lett 2004;566:110–114.
46. Walters KJ, Lech PJ, Goh AM, Wang Q, Howley PM. DNA-repair protein hHR23a alters its protein structure upon binding proteasomal subunit S5a. Proc Natl Acad Sci USA 2003;100:12694–12699.
47. Harper JW, Schulman BA. Structural complexity in ubiquitin recognition. Cell 2006;124:1133–1136.
48. Hurley JH, Lee S, Prag G. Ubiquitin-binding domains. Biochem J 2006;399:361–372.
49. Reverter D, Wu K, Erdene TG, Pan ZQ, Wilkinson KD, Lima CD. Structure of a complex between Nedd8 and the Ulp/Senp protease family member Den1. J Mol Biol 2005;345:141–151.
50. Huang TT, Nijman SMB, Mirchandani KD, Galardy PJ, Cohn MA, Haas W, Gygi SP, Ploegh HL, Bernards R, D'Andrea AD. Regulation of monoubiquitinated PCNA by DUB autocleavage. Nat Cell Biol 2006;8:341–347.

Supplementary Material
for

High incidence of ubiquitin-like domains in human ubiquitin-specific proteases

Xiao Zhu[1,2], Robert Ménard[1,2], Traian Sulea[1] *

[1] Biotechnology Research Institute, National Research Council of Canada, Montreal, Quebec H4P 2R2, Canada
[2] Department of Biochemistry, Université de Montréal, Montreal, Quebec H3C 3J7, Canada

Figure S1. Curated complete query-to-target assignment obtained by 3D-Jury consensus structure prediction for all newly identified UBL domains of human USPs (i.e., excluding those from USP14 and USP48). See the Materials and Methods section of the main text for details on 3D-Jury and its present application to human USPs.

3D-Jury scores are coded using four shades of gray (see legend). For all newly predicted UBL domains, there is at least one UBL structural template with 3D-Jury higher than 40. Only the highest of the scores corresponding to different fold recognition methods for the same structural template (i.e., the same PDB entry) is plotted. The same applies for independent structures of the same template (e.g., PDB entries 1UBI and 1UBQ of ubiquitin).

The structural templates are described at the top of the figure by their PDB codes, protein name and species. All but three of these templates belong structurally to the ubiquitin superfamily, classified in the SCOP database (http://scop.berkeley.edu/) under code d.15.1 or 54236. The only exceptions are: (i) for the USP7 sequence 792-883, where two PB1 domains were retrieved with 3D-Jury score ≥ 40 (PDB codes 1IPG and 1VD2), and (ii) for the USP47 sequence 1221-1321, where a 2Fe-2S ferrodoxin-like domain was retrieved with 3D-Jury score > 30 (PDB code 1E7P). Both the PB1 domain (SCOP code d.15.2 or 54277) and the ferrodoxin-like domain (SCOP code d.15.4 or 54292) belong to the β-grasp ubiquitin-like fold (SCOP code d.15 or 54235) and are closely related to the archetypal ubiquitin superfamily. No other folds were detected with a score higher than 30 for any of these UBL domains of human USPs. Note that some of the template protein structures identified in Figure S1 contain multiple domains; in these cases it is only their UBL domains that were identified as structural templates.

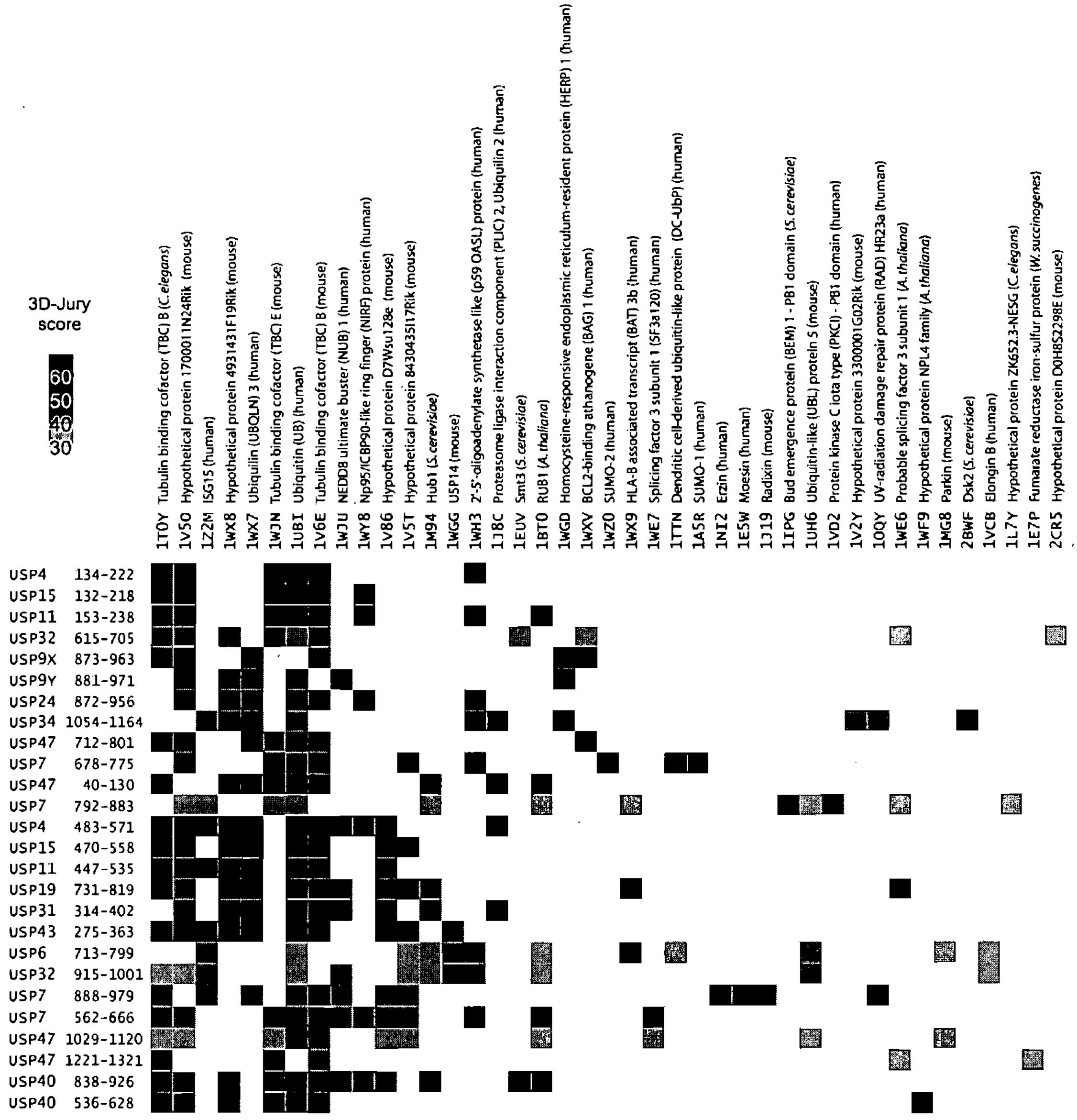
3D-Jury score
60
50
40
30
1T0Y Tubulin binding cofactor (TBC) B (*C. elegans*)
1V5O Hypothetical protein 1700011N24Rik (mouse)
1Z2M ISG15 (human)
1WX8 Hypothetical protein 4931431F19Rik (mouse)
1WX7 Ubiquilin (UBQLN) 3 (human)
1WJN Tubulin binding cofactor (TBC) E (mouse)
1UBI Ubiquitin (UB) (human)
1V6E Tubulin binding cofactor (TBC) B (mouse)
1WJU NEDD8 ultimate buster (NUB) 1 (human)
1WY8 Np95/ICBP90-like ring finger (NIRF) protein (human)
1V86 Hypothetical protein D7Wsu128e (mouse)
1V5T Hypothetical protein 8430435I17Rik (mouse)
1M94 Hub1 (*S. cerevisiae*)
1WGG USP14 (mouse)
1WH3 2'-5'-oligoadenylate synthetase like (p59 OASL) protein (human)
1J8C Proteasome ligase interaction component (PLIC) 2, Ubiquilin 2 (human)
1EUV Smt3 (*S. cerevisiae*)
1BT0 RUB1 (*A. thaliana*)
1WGD Homocysteine-responsive endoplasmic reticulum-resident protein (HERP) 1 (human)
1WXV BCL2-binding athanogene (BAG) 1 (human)
1WZ0 SUMO-2 (human)
1WX9 HLA-B associated transcript (BAT) 3b (human)
1WE7 Splicing factor 3 subunit 1 (SF3a120) (human)
1TTN Dendritic cell-derived ubiquitin-like protein (DC-UbP) (human)
1A5R SUMO-1 (human)
1NI2 Erzin (human)
1E5W Moesin (human)
1J19 Radixin (mouse)
1IPG Bud emergence protein (BEM) 1 - PB1 domain (*S. cerevisiae*)
1UH6 Ubiquitin-like (UBL) protein 5 (mouse)
1VD2 Protein kinase C iota type (PKCI) - PB1 domain (human)
1V2Y Hypothetical protein 3300001G02Rik (mouse)
1OQY UV-radiation damage repair protein (RAD) HR23a (human)
1WE6 Probable splicing factor 3 subunit 1 (*A. thaliana*)
1WF9 Hypothetical protein NPL4 family (*A. thaliana*)
1MG8 Parkin (mouse)
2BWF Dsk2 (*S. cerevisiae*)
1VCB Elongin B (human)
1L7Y Hypothetical protein ZK652.3-NESG (*C. elegans*)
1E7P Fumarate reductase iron-sulfur protein (*W. succinogenes*)
2CR5 Hypothetical protein D0H8S2298E (mouse)
USP4 134-222
USP15 132-218
USP11 153-238
USP32 615-705
USP9X 873-963
USP9Y 881-971
USP24 872-956
USP34 1054-1164
USP47 712-801
USP7 678-775
USP47 40-130
USP7 792-883
USP4 483-571
USP15 470-558
USP11 447-535
USP19 731-819
USP31 314-402
USP43 275-363
USP6 713-799
USP32 915-1001
USP7 888-979
USP7 562-666
USP47 1029-1120
USP47 1221-1321
USP40 838-926
USP40 536-628

Figure S1

Figure S2. Novel UBL domains reliably detected by consensus fold recognition in the USP family from *Saccharomyces cerevisiae* yeast. Ubiquitin-like domains were recognized in yeast Ubp12 and Ubp15 between indicated boundaries, with 3D-Jury scores above the significance cutoff of 40. Yeast Ubp6, the only yeast USP (out of 16 family members) previously annotated to contain a UBL domain, is also shown.

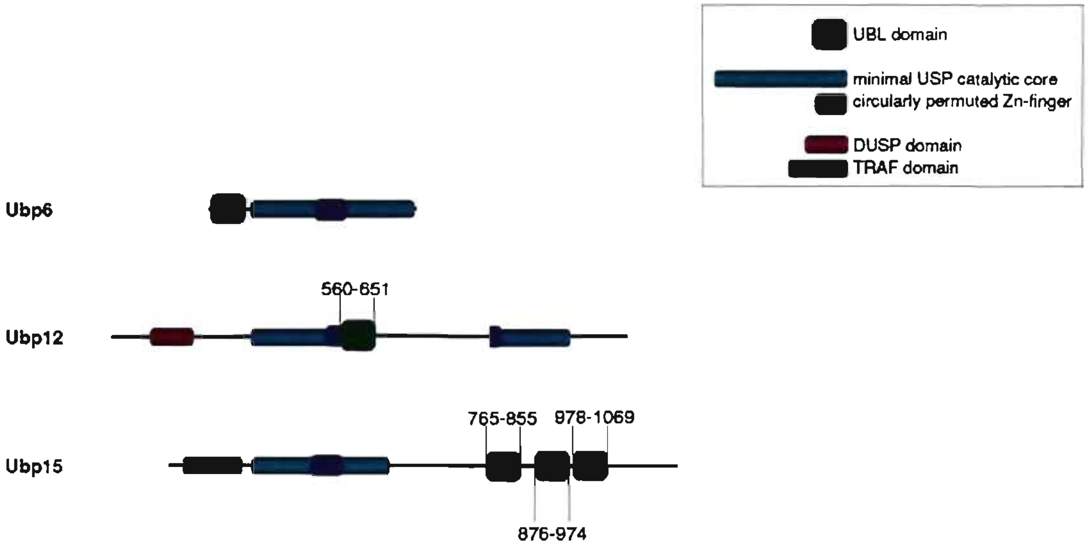

Figure S2

Table S1. Database accession numbers for the 54 human USP sequences used in this work.

Name	GenBank	SwissProt
USP1	29791616	
USP2	12804195	
USP3	55770886	
USP4	40795665	
USP5		P45974
USP6	4758564	
USP7	4507857	
USP8	41281376	
USP9X	11641425	
USP9Y	4759296	
USP10	24307889	
USP11	24234683	
USP12		O75317
USP13		Q92995
USP14	4827050	
USP15	14149627	
USP16	5454156	
USP17	551291	
USP18	15928868	
USP19	57529246	
USP20		Q9Y2K6
USP21		Q9UK80
USP22		Q9UPT9
USP24		Q9UPU5
USP25	49899220	
USP26	57284140	
USP27	113429832	
USP28	16507200	
USP29	10190742	
USP30	14249222	
USP31	42415503	
USP32	22550104	
USP33	42516567	
USP34	41056187	
USP35		Q9P2H5
USP36	35250686	
USP37	32698744	
USP38	27545313	
USP39	56550051	
USP40		Q9NVE5
USP41		Q70BM7
USP42	51094456	
USP43	40788175	
USP44	14149817	
USP45	56204652	
USP46	31377709	
USP47	71774197	
USP48	52630449	
USP49	55959716	
USP50	45267835	
USP51	41152235	
USP52	41281527	
USP53	56912182	
USP54	40255121	

Table S2. Database accession numbers for the 16 USP sequences from budding yeast.

Name	SwissProt
Ubp1	P25037
Ubp2	Q01476
Ubp3	Q01477
Ubp4	P32571
Ubp5	P39944
Ubp6	P43593
Ubp7	P40453
Ubp8	P50102
Ubp9	P39967
Ubp10	P53874
Ubp11	P36026
Ubp12	P39538
Ubp13	P38187
Ubp14	P38237
Ubp15	P50101
Ubp16	Q02863

4.1 Personal contributions

Beginning from an idea initiated by Dr. Sulea that long sequences from unstructured regions in human USPs may contain important information about the function of each USP, I have carried out the primary work of this research project and produced results as presented in the article. Under the guidance of Dr. Sulea and Dr. Ménard, I have generated and refined multiple sequence alignments, elaborated the procedure for sequence selection and analysis, and analyzed final results. I have also worked closely with Dr. Sulea, who contributed most to the production of the text, for the elaboration of the paper.

Chapter 5 Conclusion

In this research project we applied multiple sequence alignment and 3D-fold recognition to address the current limitations of identifying structural and functional features of 54 human USPs. We have identified a multitude of previously unknown domains including Zn-fingers and EF-hands. The most important discovery is the remarkably high occurrence of ubiquitin-like (UBL) domains integrated, not only at both termini of USPs but also within previously un-annotated segments embedded within the catalytic core. These key observations increase our understanding of the structural organization of USPs by presenting a new dimension in the structural and functional complexity of these enzymes.

In addition to the presence of integral UBL domains in human USPs, we have uncovered many other structurally and functionally diverse regions with significant confidence level. For example, further analyses of the sequences in the upstream and downstream regions of newly identified UBL domains (in USP9X, USP9Y, USP24, and USP34) revealed long stretches of armadillo repeats, which may form an all-alpha helix bundle much like nuclear transporters, and interact with nuclear receptors and transporters (Figure 3). We have also identified an additional CS domain immediately upstream of the known domain (PDB: 1WH0). The occurrence of α-helical structural arrangement similar to nuclear receptors and beta-catenin, and generally known as armadillo (ARM) repeats, was found in many USPs (Figure 3). Among them, USPs 9X, 9Y, 24, and 34 harbor an UBL domain at the heart of their long ARM stretches. We believe that these all-alpha regions may interact with nuclear receptors and transporters and also modulate Wnt signaling and cell adhesion (Figure 3). Finally, a zinc-finger UBP domain was identified in USP39 where it may play a role in Ub binding and hydrolysis (Figure 3).

Our findings present leverage for future investigations on the role of these newly discovered domains in substrate specificity and subcellular localization of USPs as well as their physiological functions in human diseases. While further biochemical or biophysical analyses will be required to validate the accuracy of our predictions, our present results provide solid indication of the utility of structure-based functional prediction by associating sequences lacking any detectable homologies. The consensus method for fold prediction and recognition is a relatively new concept which may require time to be fully acknowledged. We believe that this type of approach should be integrated routinely into first degree functional assignments where sequence-based comparisons would fail.

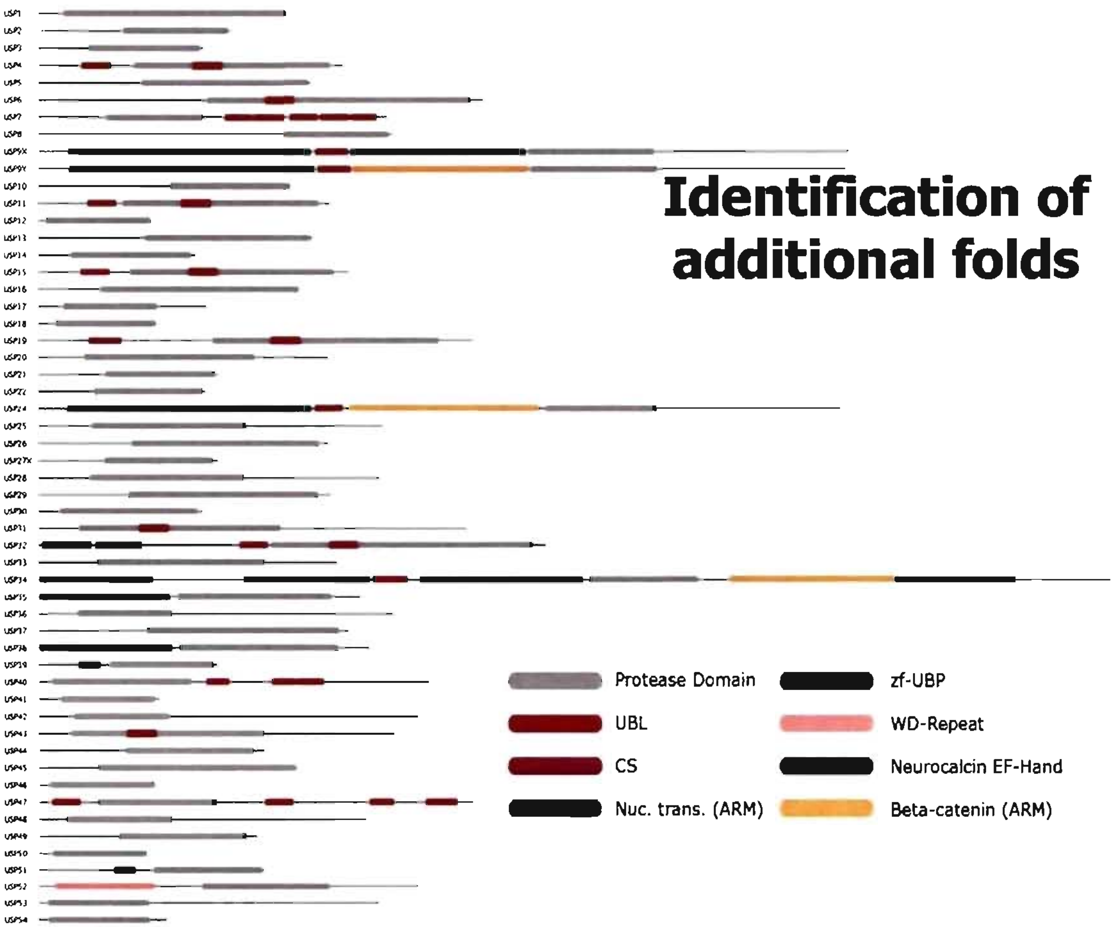

Figure 3: Identification of novel folds within the USP family with catalytic core represented as protease domain in gray. In addition to the frequent occurrence of UBL domains in the 54 USPs analyzed, nuclear transporter-like and be0074a-catenin-like armadillo repeats were identified at both N- and C-terminal regions of USPs 9X, 9Y, 24, 34, 35, and 38. Zinc-finger UBPs were found in USPs 39 and 51. A novel CS domain was identified in USP19. The all-beta sheet WD-repeat region was found in USP52. Two EF-hands with close resemblance to those in neurocalcin were identified in USP32.

Reference List

Anderson,C., Crimmins,S., Wilson,J.A., Korbel,G.A., Ploegh,H.L., and Wilson,S.M. 2005. Loss of Usp14 results in reduced levels of ubiquitin in ataxia mice. *J. Neurochem.* **95**:724-731.

Avvakumov,G.V., Walker,J.R., Xue,S., Finerty,P.J., Jr., Mackenzie,F., Newman,E.M., and Dhe-Paganon,S. 2006. Amino-terminal dimerization, NRDP1-rhodanese interaction, and inhibited catalytic domain conformation of the ubiquitin-specific protease 8 (USP8). *J. Biol. Chem.* **281**:38061-38070.

Bairoch,A., Boeckmann,B., Ferro,S., and Gasteiger,E. 2004. Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.* **5**:39-55.

Blanchette,P., Gilchrist,C.A., Baker,R.T., and Gray,D.A. 2001. Association of UNP, a ubiquitin-specific protease, with the pocket proteins pRb, p107 and p130. *Oncogene* **20**:5533-5537.

Borodovsky,A., Kessler,B.M., Casagrande,R., Overkleeft,H.S., Wilkinson,K.D., and Ploegh,H.L. 2001. A novel active site-directed probe specific for deubiquitylating enzymes reveals proteasome association of USP14. *EMBO J.* **20**:5187-5196.

Bourne,P.E. 2003. CASP and CAFASP experiments and their findings. *Methods Biochem. Anal.* **44**:501-507.

Brenner,S.E., Chothia,C., and Hubbard,T.J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. U. S. A* **95**:6073-6078.

Crimmins,S., Jin,Y., Wheeler,C., Huffman,A.K., Chapman,C., Dobrunz,L.E., Levey,A., Roth,K.A., Wilson,J.A., and Wilson,S.M. 2006. Transgenic rescue of ataxia mice with neuronal-specific expression of ubiquitin-specific protease 14. *J. Neurosci.* **26**:11423-11431.

Cummins,J.M., Rago,C., Kohli,M., Kinzler,K.W., Lengauer,C., and Vogelstein,B. 2004. Tumour suppression: disruption of HAUSP gene stabilizes p53. *Nature* **428**:1.

D'Amato,C.J. and Hicks,S.P. 1965. Neuropathologic alterations in the ataxia (paralytic) mouse. *Arch. Pathol.* **80**:604-612.

Daviet,L. and Colland,F. 2007. Targeting ubiquitin specific proteases for drug discovery. *Biochimie.*

Everett,R.D., Meredith,M., Orr,A., Cross,A., Kathoria,M., and Parkinson,J. 1997. A novel ubiquitin-specific protease is dynamically associated with the PML nuclear domain and binds to a herpesvirus regulatory protein. *EMBO J.* **16**:1519-1530.

Gewies,A. and Grimm,S. 2003. UBP41 is a proapoptotic ubiquitin-specific protease. *Cancer Res.* **63**:682-688.

Ginalski,K., Elofsson,A., Fischer,D., and Rychlewski,L. 2003. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics.* **19**:1015-1018.

Ginalski,K., Grishin,N.V., Godzik,A., and Rychlewski,L. 2005. Practical lessons from protein structure prediction. *Nucleic Acids Res.* **33**:1874-1891.

Gong,L., Kamitani,T., Millas,S., and Yeh,E.T. 2000. Identification of a novel isopeptidase with dual specificity for ubiquitin- and NEDD8-conjugated proteins. *J. Biol. Chem.* **275**:14212-14216.

Graner,E., Tang,D., Rossi,S., Baron,A., Migita,T., Weinstein,L.J., Lechpammer,M., Huesken,D., Zimmermann,J., Signoretti,S., and Loda,M. 2004. The isopeptidase USP2a regulates the stability of fatty acid synthase in prostate cancer. *Cancer Cell* **5**:253-261.

Hetfeld,B.K., Helfrich,A., Kapelari,B., Scheel,H., Hofmann,K., Guterman,A., Glickman,M., Schade,R., Kloetzel,P.M., and Dubiel,W. 2005. The zinc finger of the CSN-associated deubiquitinating enzyme USP15 is essential to rescue the E3 ligase Rbx1. *Curr. Biol.* **15**:1217-1221.

Hicke,L. and Dunn,R. 2003. Regulation of membrane protein transport by ubiquitin and ubiquitin-binding proteins. *Annu. Rev. Cell Dev. Biol.* **19**:141-172.

Holler,D. and Dikic,I. 2004. Receptor endocytosis via ubiquitin-dependent and -independent pathways. *Biochem. Pharmacol.* **67**:1013-1017.

Hu,M., Gu,L., Li,M., Jeffrey,P.D., Gu,W., and Shi,Y. 2006. Structural basis of competitive recognition of p53 and MDM2 by HAUSP/USP7: implications for the regulation of the p53-MDM2 pathway. *PLoS. Biol.* **4**:e27.

Hu,M., Li,P., Song,L., Jeffrey,P.D., Chenova,T.A., Wilkinson,K.D., Cohen,R.E., and Shi,Y. 2005. Structure and mechanisms of the proteasome-associated deubiquitinating enzyme USP14. *EMBO J.* **24**:3747-3756.

Jenuth,J.P. 2000. The NCBI. Publicly available tools and resources on the Web. *Methods Mol. Biol.* **132**:301-312.

Kerscher,O., Felberbaum,R., and Hochstrasser,M. 2006. Modification of proteins by ubiquitin and ubiquitin-like proteins. *Annu. Rev. Cell Dev. Biol.* **22**:159-180.

Kiel,C. and Serrano,L. 2006. The ubiquitin domain superfold: structure-based sequence alignments and characterization of binding epitopes. *J. Mol. Biol.* **355**:821-844.

Kim,J.H., Park,K.C., Chung,S.S., Bang,O., and Chung,C.H. 2003. Deubiquitinating enzymes as cellular regulators. *J. Biochem. (Tokyo)* **134**:9-18.

Krogh,A., Mian,I.S., and Haussler,D. 1994. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res.* **22**:4768-4778.

Li,M., Brooks,C.L., Kon,N., and Gu,W. 2004. A dynamic role of HAUSP in the p53-Mdm2 pathway. *Mol. Cell* **13**:879-886.

Loayza,D. and Michaelis,S. 1998. Role for the ubiquitin-proteasome system in the vacuolar degradation of Ste6p, the a-factor transporter in Saccharomyces cerevisiae. *Mol. Cell Biol.* **18**:779-789.

Lundstrom,J., Rychlewski,L., Bujnicki,J., and Elofsson,A. 2001. Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.* **10**:2354-2362.

Malakhov,M.P., Malakhova,O.A., Kim,K.I., Ritchie,K.J., and Zhang,D.E. 2002. UBP43 (USP18) specifically removes ISG15 from conjugated proteins. *J. Biol. Chem.* **277**:9976-9981.

Martin,A.C., Orengo,C.A., Hutchinson,E.G., Jones,S., Karmirantzou,M., Laskowski,R.A., Mitchell,J.B., Taroni,C., and Thornton,J.M. 1998. Protein folds and functions. *Structure.* **6**:875-884.

Milojevic,T., Reiterer,V., Stefan,E., Korkhov,V.M., Dorostkar,M.M., Ducza,E., Ogris,E., Boehm,S., Freissmuth,M., and Nanoff,C. 2006. The ubiquitin-specific protease Usp4 regulates the cell surface level of the A2A receptor. *Mol. Pharmacol.* **69**:1083-1094.

Murzin,A.G., Brenner,S.E., Hubbard,T., and Chothia,C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**:536-540.

Naviglio,S., Mattecucci,C., Matoskova,B., Nagase,T., Nomura,N., Di Fiore,P.P., and Draetta,G.F. 1998. UBPY: a growth-regulated human ubiquitin isopeptidase. *EMBO J.* **17**:3241-3250.

Nijman,S.M., Luna-Vargas,M.P., Velds,A., Brummelkamp,T.R., Dirac,A.M., Sixma,T.K., and Bernards,R. 2005. A genomic and functional inventory of deubiquitinating enzymes. *Cell* **123**:773-786.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B., and Thornton,J.M. 1997. CATH--a hierarchic classification of protein domain structures. *Structure.* **5**:1093-1108.

Peschard,P. and Park,M. 2003. Escape from Cbl-mediated downregulation: a recurrent theme for oncogenic deregulation of receptor tyrosine kinases. *Cancer Cell* **3**:519-523.

Pickart,C.M. and Eddins,M.J. 2004. Ubiquitin: structures, functions, mechanisms. *Biochim. Biophys. Acta* **1695**:55-72.

Priolo,C., Tang,D., Brahamandan,M., Benassi,B., Sicinska,E., Ogino,S., Farsetti,A., Porrello,A., Finn,S., Zimmermann,J., Febbo,P., and Loda,M. 2006. The isopeptidase USP2a protects human prostate cancer from apoptosis. *Cancer Res.* **66**:8625-8632.

Qiu,X.B., Markant,S.L., Yuan,J., and Goldberg,A.L. 2004. Nrdp1-mediated degradation of the gigantic IAP, BRUCE, is a novel pathway for triggering apoptosis. *EMBO J.* **23**:800-810.

Quesada,V., az-Perales,A., Gutierrez-Fernandez,A., Garabaya,C., Cal,S., and Lopez-Otin,C. 2004. Cloning and enzymatic analysis of 22 novel human ubiquitin-specific proteases. *Biochem. Biophys. Res. Commun.* **314**:54-62.

Rychlewski,L. and Fischer,D. 2005. LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci.* **14**:240-245.

Rychlewski,L., Fischer,D., and Elofsson,A. 2003. LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins* **53 Suppl 6**:542-547.

Sacks,W.R. and Schaffer,P.A. 1987. Deletion mutants in the gene encoding the herpes simplex virus type 1 immediate-early protein ICP0 exhibit impaired growth in cell culture. *J. Virol.* **61**:829-839.

Scheffner,M., Takahashi,T., Huibregtse,J.M., Minna,J.D., and Howley,P.M. 1992. Interaction of the human papillomavirus type 16 E6 oncoprotein with wild-type and mutant human p53 proteins. *J. Virol.* **66**:5100-5105.

Scheffner,M., Werness,B.A., Huibregtse,J.M., Levine,A.J., and Howley,P.M. 1990. The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell* **63**:1129-1136.

Soetens,O., De Craene,J.O., and Andre,B. 2001. Ubiquitin is required for sorting to the vacuole of the yeast general amino acid permease, Gap1. *J. Biol. Chem.* **276**:43949-43957.

Stevenson,L.F., Sparks,A., lende-Vega,N., Xirodimas,D.P., Lane,D.P., and Saville,M.K. 2007. The deubiquitinating enzyme USP2a regulates the p53 pathway by targeting Mdm2. *EMBO J.* **26**:976-986.

Sun,C., Skaletsky,H., Birren,B., Devon,K., Tang,Z., Silber,S., Oates,R., and Page,D.C. 1999. An azoospermic man with a de novo point mutation in the Y-chromosomal gene USP9Y. *Nat. Genet.* **23**:429-432.

Wada,K. and Kamitani,T. 2006. UnpEL/Usp4 is ubiquitinated by Ro52 and deubiquitinated by itself. *Biochem. Biophys. Res. Commun.* **342**:253-258.

Wallner,B., Fang,H., and Elofsson,A. 2003. Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins* **53 Suppl 6**:534-541.

Walters,K.J., Goh,A.M., Wang,Q., Wagner,G., and Howley,P.M. 2004. Ubiquitin family proteins and their relationship to the proteasome: a structural perspective. *Biochim. Biophys. Acta* **1695**:73-87.

Wang,G., Jin,Y., and Dunbrack,R.L., Jr. 2005. Assessment of fold recognition predictions in CASP6. *Proteins* **61 Suppl 7**:46-66.

Welchman,R.L., Gordon,C., and Mayer,R.J. 2005. Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nat. Rev. Mol. Cell Biol.* **6**:599-609.

Wilkinson,K.D. 1995. Roles of ubiquitinylation in proteolysis and cellular regulation. *Annu. Rev. Nutr.* **15**:161-189.

Wu,X., Yen,L., Irwin,L., Sweeney,C., and Carraway,K.L., III 2004. Stabilization of the E3 ubiquitin ligase Nrdp1 by the deubiquitinating enzyme USP8. *Mol. Cell Biol.* **24**:7748-7757.