# Mining the TRAF6/P62 Interactome for Preferred Substrates and Target Ubiquitination Sites: Developing a "Code Hypothesis"

by

Trafina Jadhav

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 9, 2010

Keywords: TRAF6, p62, ubiquitination,
computational, TrkA, Lysine

Approved by

Marie W. Wooten, Chair, Professor of Biological Sciences
Scott R. Santos, Professor of Biological Sciences
Narendra Singh, Professor of Biological Sciences
Michael C. Wooten, Professor of Biological Sciences

# Abstract

Ubiquitination is the second most common protein modification studied in terms of biochemistry and cell physiology. It plays a central regulatory role in number of eukaryotic cellular and molecular processes. This three step process of concerted action of the E1-E2-E3 enzymes produces an ubiquitinated protein. How E3 ligases select substrates and achieve selectivity at a Lysine residue remains unsolved. I undertook studies to identify both ubiquitin and SUMO (small ubiquitin-related modifier) substrates with the goal of understanding how Lysine selectivity is achieved in these two processes. Although distinct from ubiquitination, SUMOylation pathway draws many parallels with it. Based upon recent findings, I present a model that explains how an individual ubiquitin ligase may target specific Lysine residue(s) with the co-operation from a scaffold protein, p62. Tumor necrosis factor receptor-associated factor 6 (TRAF6) is an ubiquitin ligase that regulates a diverse array of physiological processes *via* forming Lys-63 linked polyubiquitin chains. Described here is a new approach to predict ubiquitinated substrates of TRAF6/p62 complex. Interactome knowledge was used to predict potential TRAF6 substrates. Observations showed that there was low linear conservation of a single consensus motif at predicted ubiquitinated sites. However, a substantial structural and sequence conservation was observed across mammalian species for a novel ubiquitination defined as [–(hydrophobic)–k–(hydrophobic)–x–x–(hydrophobic)–(polar)–(hydrophobic) –(polar)–(hydrophobic)]. These findings revealed that the identified target sites have structural preferences and depend on accessibility within the protein molecule.

# Acknowledgements

First and above all, I praise God, the almighty for providing me this opportunity and granting me the fortitude that has carried me through to proceed successfully. My doctoral investigations have been conducted under the auspices of many individuals. I find myself indebted to Dr. Marie W. Wooten for her sterling mentorship, insights in signal transduction studies, and for instilling rigor for research and critical thinking. I am also extremely grateful to Dr. Michael C. Wooten for his technical guidance with the computer programming and for his approachable demeanor. Committee members Drs Scott Santos and Narendra Singh deserved a special mention too; as they dedicated their time and energy into reviewing manuscripts and making sure I get the most out of not only my dissertation research, but also my career. I also wish to thank Drs Jacek Wower and Bertram Zinner for providing constructive suggestions concerning this work. I am grateful to Ms Sandra Riddle, Dr. Mary Mendonca, and Dr. Jack Feminella for providing support at the department of Biological Sciences to carry out the studies. I also want to thank Wooten Lab members Lamar Seibenhener, Yifeng Du, and Geetha Thangiah. I am especially grateful to Diane and Len Vining for helping me get through the difficult times, and for their friendship and family-like unwavering support when I needed it the most. Finally, I express gratitude to my dear friends, for enriching my life with indelible memories and camaraderie; I could not have succeeded without the invaluable support from Ryan Schumacher, Nirmal Andrews, Devi Anamalai, Tom Williams, Yogesh Kondareddy, Dr. Larissa Parsley, Diana Amrapalli, Bujung Hadi, Paul Dawson, Yifeng Du, Michael Madan and Saishree. Of course no acknowledgments

would be complete without giving thanks to my parents and family for understanding my desire to pursue doctorate in United States. They all have in some ways have instilled many admirable qualities in me and have taught me about hard work and self-respect, about persistence and about how to be independent and develop a strong character.

## Table of Contents

Chapter 3. Computational search for preferred TRAF6/p62 ubiquitination sites: a test of the "code-hypothesis"

Chapter 4. Summary and future directions

## List of Figures

# List of Tables

## List of Abbreviations

| | |
|---|---|
| AID | Atypical PKC-interaction domain |
| ATG8 | Autophagy associated protein 8 |
| ATG12 | Autophagy associated protein 12 |
| CHIP | C-terminus of Hsc70-interacting protein |
| DUB | De-ubiquitinating enzyme |
| HECT | Homologous to E6-AP C Terminus |
| HIF1 | Hypoxia inducible factor-1 |
| FUB1 | FBR-MuSV associated ubiquitously expressed gene |
| KO | Knock-out |
| HEK | Human embryonic kidney |
| IP | Immunoprecipitation |
| ISG15 | Interferon-stimulated gene 15 |
| MBP | Myelin basic protein |
| NEDD8 | Neural precursor cell expressed, developmentally down-regulated 8 |
| NRIF | Neurotrophin receptor interacting factor |
| NTRK2 | Neurotrophic tyrosine receptor kinase 2 |
| NTRK3 | Neurotrophic tyrosine receptor kinase 3 |
| PAGE | Polyacrylamide gel electrophoresis |

| | |
|---|---|
| PB1 | Phox and Bem1 |
| PBS | Phosphate buffer saline |
| PEST | Proline Glutamate Serine Threonine |
| PKC | Protein kinase C |
| RING | Really Interesting New Gene |
| SCF | Skp1-Cullin-F-box protein |
| SDS | Sodium dodecyl sulfate |
| SUMO | Small ubiquitin-related modifier |
| TRAF6 | Tumor necrosis factor receptor associated factor 6 |
| TrkA | Tropomyosin receptor kinase A |
| TrkB | Tropomyosin receptor kinase B |
| TrkC | Tropomyosin receptor kinase C |
| Ub | Ubiquitin |
| UBA | Ubiquitin-associated domain |
| UBD | Ubiquitin-binding domains |
| UBL | Ubiquitin-like protein |
| UBL5 | Ubiquitin-like 5 |
| URM1 | Ubiquitin-related modifier 1 |
| VHL | Von Hippel Lindau protein |
| WB | Western blot |
| WT | Wild-type |
| ZIP | Zeta protein knase C interacting protein |
| ZZ | ZZ-type Zinc finger domain |

# CHAPTER 1. LITERATURE REVIEW

# DEFINING AN EMBEDDED CODE FOR PROTEIN UBIQUITINATION

## ABSTRACT

It has been more than 30 years since the initial report of the discovery of ubiquitin as an 8.5 kDa protein of unknown function expressed universally in living cells. And still, protein modification by covalent conjugation of the ubiquitin molecule is one of the most dynamic posttranslational modifications studied in terms of biochemistry and cell physiology. Ubiquitination plays a central regulatory role in number of eukaryotic cellular processes such as receptor endocytosis, growth-factor signaling, cell-cycle control, transcription, DNA repair, gene silencing, and stress response. Ubiquitin conjugation is a three step concerted action of the E1-E2-E3 enzymes that produces a modified protein. In this review I investigate studies undertaken to identify both ubiquitin and SUMO (small ubiquitin-related modifier) substrates with the goal of understanding how Lysine selectivity is achieved. The SUMOylation pathway though distinct from that of ubiquitination, draws many parallels. Based upon the recent findings, I present a model to explain how an individual ubiquitin ligase may target specific Lysine residue(s) with the co-operation from a scaffold protein.

* Published in J Proteomics Bioinform. 2009 Jul 24; 2:316.

## INTRODUCTION

Ubiquitination was originally described as a mechanism by which cells disposed of short-lived, damaged or abnormal proteins. However, its involvement in diverse cellular processes is coming to light and considered to rival phosphorylation. Ubiquitination is an ATP-requiring process and at the center of this modification is ubiquitin a 76-amino acid (~9 kDa) protein (Figure 1), which is highly conserved across eukaryotes and is synthesized as a fusion protein either to itself or to one of two ribosomal proteins **(**Schlesinger et al., 1987). Conjugation involves attachment of C-terminal glycine of ubiquitin (Ub) to the ε-amino group in Lysine residues of the targeted protein. The conserved conjugation reaction is achieved by sequential actions of three enzymes (Hershko et al., 1998). The reaction commences with the formation of a thiol-ester linkage between the glycine residue at the C terminus of Ub and the active cysteine (Cys) residue of the first enzyme of the system, Ub activating enzyme (commonly referred to as E1). The ubiquitin molecule is then subsequently transferred to the cysteinyl group of the second enzyme called Ub-conjugating enzyme (E2). Lastly, through the action of an Ub ligase (E3), ubiquitin and the marked substrate are linked together *via* an amide (isopeptide) bond. This ability of an E3 to recognize and bind both the target substrate and the Ub-E2 enzyme suggests this enzyme provides specificity to the Ub reaction. At this point, the ubiquitination reaction may result in the addition of a single Ub molecule to a single target site, mono-ubiquitination (Figure 2). Alternatively, ubiquitination may result in the addition of single molecules of ubiquitin to other Lys in the target protein giving rise to multi-ubiquitination. After the initial ubiquitin is

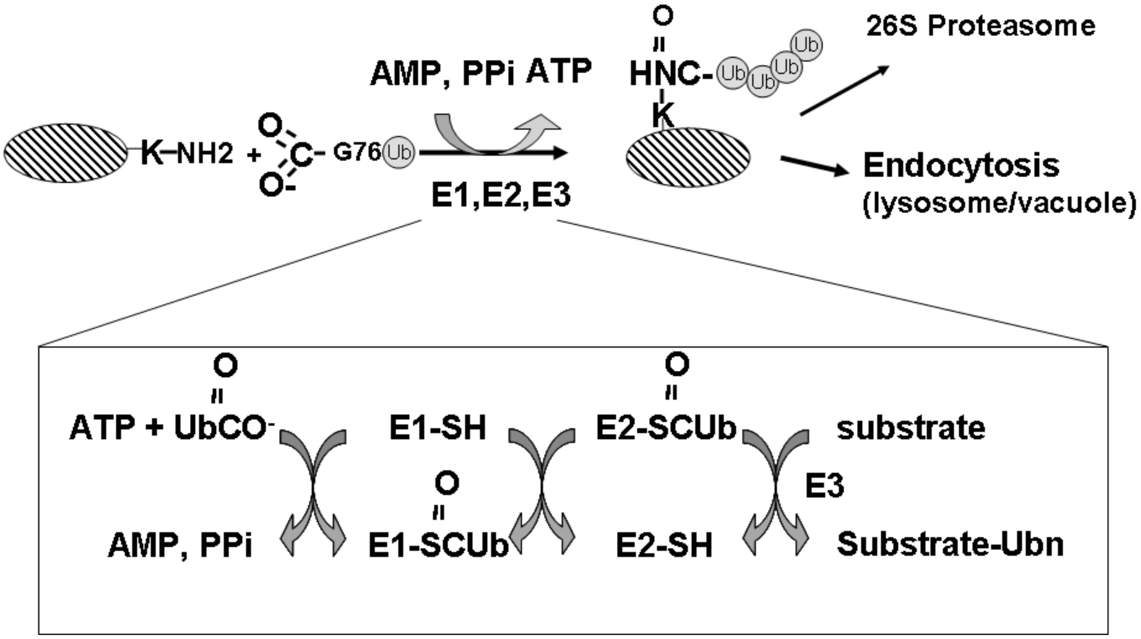

26S Proteasome
AMP, PPi
ATP
O
HNC-
Ub
Ub
Ub
Ub
K
K-NH2
+
O
C
O-
G76
Ub
E1,E2,E3
Endocytosis
(lysosome/vacuole)
O
ATP + UbCO-
E1-SH
O
E2-SCUb
substrate
O
E3
AMP, PPi
E1-SCUb
E2-SH
Substrate-Ubn

**Figure 1.** Ubiquitination reaction. The protein substrate is ubiquitinated in a reaction involving three types of ubiquitinating enzymes: the ubiquitin activating protein E1, an ubiquitin carrier protein E2, and an ubiquitin-protein ligase E3. Following addition of a single ubiquitin molecule to a protein substrate (monoubiquitination), further ubiquitin molecules can be added to the first, yielding a polyubiquitin chain. The fate of the protein depends on the type of ubiquitin chain formed on the protein substrate.

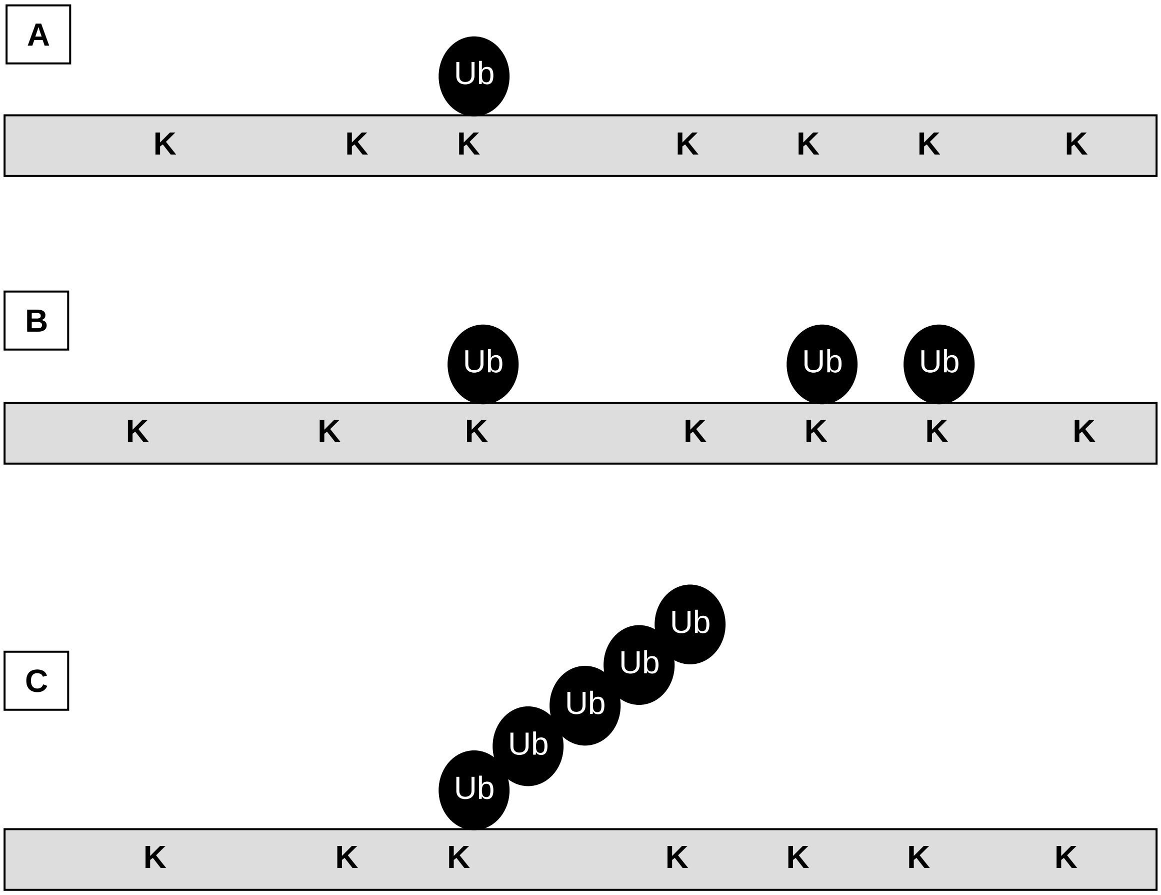

A
Ub
K
K
K
K
K
K
K
B
Ub
Ub
Ub
K
K
K
K
K
K
K
C
Ub
Ub
Ub
Ub
Ub
K
K
K
K
K
K
K

**Figure 2.** Ubiquitin modifications**. A.** *Mono-ubiquitination* is involved in transcription, histone function, endocytosis and membrane trafficking. **B.** *Multi-monoubiquitination* is involved in protein regulation. **C.** *Polyubiquitination* is involved in signal transduction, endocytosis, DNA repair, stress response, and targeting proteins to the proteasome.

conjugated to a substrate, it can also be conjugated to another molecule of ubiquitin through one of its seven Lysines. An isopeptide bond is formed between Gly76 of one ubiquitin to the ε-$NH_2$ group of one of the seven potential Lysines (K6, K11, K27, K29, K33, K48 or K63) of the preceding ubiquitin, giving rise to many different types of poly-ubiquitinated proteins (Adhikari and Chen, 2009)**.** These poly-ubiquitin chains can vary in length with respect to the number of ubiquitin molecules, resulting in different topologies and, ultimately different functional consequences. For example, Lys48-linked polyubiquitination primes proteins for proteolytic destruction by the proteasome (Chau et al., 1989), whereas Lys63-linked polyubiquitination plays a key role in regulating processes such as DNA repair (Spence et al., 1995; Hofmann and Pickart, 1999), stress responses (Arnason and Ellison, 1994), signal transduction (Sun and Chen, 2004; Mukhopadhyay and Riezman, 2007), and intracellular trafficking of membrane proteins (Hicke, 1999; Geetha et al., 2005; Mukhopadhyay and Riezman, 2007).

Proteins tagged with ubiquitin are most often destined for degradation by the proteasome. Recent studies reveal that all non-K63 linkages may target proteins for degradation (Xu et al., 2009). However this is still a matter of debate since K63-chains have also been shown to serve as a targeting signal for the 26S proteasome (Seibenhener et al., 2004; Saeki et al., 2009). Both, mono-ubiquitination and poly-ubiquitination also possess non-proteasomal regulatory functions like targeting proteins to nucleus, cytoskeleton and endocytic machinery, or modulating enzymatic activity and protein-protein interactions (Hershko et al., 1998; Pickart, 2001). Recent reports have indicated non Lysine moieties can serve as ubiquitin acceptor sites. Ubiquitination occurring at noncanonical site —the N terminus— has been reported for transcription factor

MyoD, the latent membrane protein-1 of Epstein-Barr virus, and p21, lead to proteasome-mediated degradation (Aviel et al., 2000; Breitschopf et al., 1998; Bloom et al., 2003). Moreover, studies have shown the cysteine residue is required for ubiquitination of major histocompatibility complex class I proteins by the viral E3 ligases (Cadwell and Coscoy, 2005). Like other posttranslational modifications (e.g. phosphorylation) ubiquitination is highly regulated and reversible process. It is controlled by the opposing activities of the E3 protein ubiquitin ligases which attach Ub molecules covalently to target proteins and de-ubiquitinating enzymes (DUBs) which remove the ubiquitin from target proteins (Wilkinson et al., 1997). Reversible covalent modification allows cells to rapidly and efficiently convey signals across different sub-cellular locations. It has been predicted that the human genome encodes three Ub-protein E1 enzymes, about fifty Ub-protein E2 conjugating complexes, over 600 ubiquitin ligases and about 100 DUBs (Kaiser and Huang, 2005).

Lysine residues are a target for diverse posttranslational modification enzymes which either attach methyl, acetyl, hydroxyl, ubiquitin or SUMO moieties to it. Except for hydroxylation, all of these attachments are reversible. In addition to ubiquitin, several ubiquitin-like proteins (Ubls) can also be conjugated to alter the function of the substrate proteins at Lysine residues. These small molecular modifiers include NEDD8 (neural precursor cell expressed, developmentally down-regulated 8), ISG15 (interferon-stimulated gene 15), FAT10, FUB1 (FBR-MuSV associated ubiquitously expressed gene), UBL5 (ubiquitin-like 5), URM1 (ubiquitin-related modifier 1), ATG8 (autophagy associated protein 8), ATG12 (autophagy associated protein 12), and three SUMO isoforms to which ubiquitin bears much resemblance

(Kerscher et al., 2006). However, modification of these Ubls requires their own unique combinations of E1, E2 and E3 and addition of these tags to the target protein likely serves a different function compared ubiquitination. These protein tags have been implicated in numerous cellular activities including DNA synthesis and repair, transcription, translation, organelle biogenesis, cell cycle control, signal transduction, protein quality control in the endoplasmic reticulum, immune system etc (Kerscher et al., 2006). These different Ubls are activated and conjugated to their substrates by a process very similar to the biochemical reactions of ubiquitination. All the structurally characterized Ubls share the ubiquitin or β-grasp fold, even when their primary sequences have little similarity (Kerscher et al., 2006).

Like several other posttranslational modifications, ubiquitination changes the molecular conformation of a protein, thereby influencing protein-protein interactions. Ubiquitin modification is known to alter protein localization, activity and/or stability through interaction with various proteins. These modifications on the target protein (either through monoubiquitination or polyubiquitination) act as attachment sites for proteins with ubiquitin-binding domains (UBDs) (Bertolaet et al., 2001; Wilkinson et al., 2001). The first UBD was characterized in a proteasome subunit, the S5A/RPN10 protein11. Similarity searches of a short sequence of S5a bound to ubiquitin led to the identification of a sequence pattern known as the ubiquitin-interacting motif (UIM) (Hofmann and Falquet, 2001). The ubiquitin-associated domain (UBA) was identified as a common sequence motif present in multiple proteins participating in ubiquitin-dependent signaling pathways (Hofmann and Bucher, 1996). Of the total sixteen UBDs reported to date, discovery of UIM and UBA domains, was the most important as it propelled the study of ubiquitination. Both UBA and UIM are known to bind poly- and mono- ubiquitin chains. The other ubiquitin-binding domains include a diverse family of structurally dissimilar protein domains, such as MIU, DUIM, CUE, GAT, NZF, A20 ZnF, UBP ZnF, UBZ, Ubc, Uev, UBM, GLUE, Jab1/MPN, and PFU (Hurley et al., 2006). Of these, many UBA-containing proteins are reported to bind polyubiquitin chains, some serve as shuttling factors for delivery of ubiquitinated proteins to the proteasome (e.g. hHR23A, p62 and Dsk2) (Seibenhener et al., 2004). This function is thought to be achieved by binding of the UBA domain to the ubiquitinated substrates, while simultaneously interacting with the proteasome through another domain (like Ubl domain) (Seibenhener et al., 2004).

Ubiquitin-protein ligases (E3) are the last (but likely the most important) components in the ubiquitin conjugation system because they play an important role in controlling target specificity. The E3s recruit target proteins, position them for optimal transfer of the Ub moiety from the E2 to a Lysine residue in the target protein, and initiate the conjugation. Ubiquitin E3 ligases can be either monomeric proteins or multimeric complexes with the most common type of Ub ligases grouped into two classes depending on their modular architecture and catalytic mechanism. Typically E3s containing a HECT domain (Homologous to E6-AP C Terminus) forms a direct thioester bond with ubiquitin. Their approximately 350 amino acid HECT domains contain a conserved Cys residue that participates in the direct transfer of activated ubiquitin from the E2 to a target protein (Hershko et al., 1998; Pickart, 2001). On the other hand, RING (Really Interesting New Gene) finger domain ligase consists of Cys and His residues that coordinate two $Zn^{++}$ ions. The globular architecture of the domain primarily functions as a scaffold for the interaction of E2s with their target proteins (Hershko et al., 1998; Pickart, 2001). These ligases require a structural and/or catalytic motif that facilitates ubiquitination without directly forming a bond with ubiquitin. RING finger domain containing E3s comprise the largest ligase family, and contain both monomeric and multimeric ubiquitin ligases. There are three types of multisubunit E3s —SCF (Skp1-Cullin-F-box protein), the APC, and the VHL (von Hippel Lindau protein) E3(s) — where a small RING finger protein is an essential component. A lesser known family of Ub E3 ligases includes an E2-binding domain called the U-box adaptor E3 ligases. The U-box ligase was first identified in yeast Ufd2 acting as an accessory protein (E4) promoting polyubiquitination of another E3's substrate (Kuhlbrodt et al., 2005). Bioinformatics studies placed them under conventional RING E3 ligases, as the U-box ligases adopt a RING domain-like conformation *via* electrostatic interactions (Aravind and Koonin et al., 2000). Genome-wide annotation of the human E3 superfamily genes (Li et al., 2008) had revealed the number of putative E3 genes, 617, to be greater than the number of human genes for protein kinases, 518, suggesting the extent of biological targets of ubiquitination.

## SUBSTRATE SELECTION FOR UBIQUITINATION

One salient question is what determines whether or not a protein is tagged by Ub? While as of yet this cannot fully be answered, recent research has uncovered some interesting clues. It has been proposed that proteins contain an “embedded code” that is recognized by the Ub machinery (Figure 3). For example, E3 ubiquitin ligases recognize their corresponding protein substrates *via* a variety of structural determinants, including primary sequence, post-translational modifications and protein folding state. Herein, I consider some of the other examples discovered thus far for directing target specificity.

### *The N-end rule*

There exists a correlation between the half-life of a protein and its N-terminal residue (Bachmair et al., 1986). The stability of a protein is dependent on the nature of its N-terminal amino acid residues, which are classified either as stabilizing or destabilizing residues. Proteins with N-terminal Met, Ser, Ala, Thr, Val, or Gly are known to have half-lives greater than 20 hours. In contrast, proteins with N-terminal Phe, Leu, Asp, Lys, or Arg have half-lives of 3 min or less. The N-end rule pathway is a proteolytic pathway targeting proteins for degradation through destabilizing N-terminal residues (N-degrons). An N-degron consists of a protein's destabilizing N-terminal residue and an internal Lys residue. E3 Ub ligases that recognize these N-degrons are called N-recognins, which share a ≈70-residue motif called the UBR box. UBR1 (also known as E3α) is the recognition component of the N-end rule pathway that binds to a destabilizing N-terminal residue of a substrate protein and participates in the formation of a substrate-linked polyubiquitin chain. Mutations in human Ubr1 have been associated with the Johansson–Blizzard Syndrome (JBS), which includes mental retardation, physical malformations and pancreatic dysfunction (Zenker et al., 2005). The N-end rule has a hierarchical structure in which primary, secondary and tertiary destabilizing N-terminal residues participate differentially based on their requirements for enzymatic modification. Recent studies have shown that though

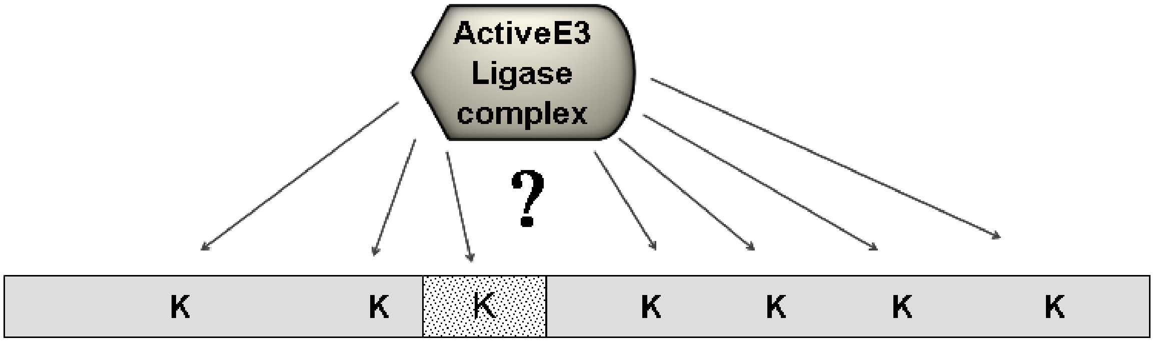

ActiveE3
Ligase
complex
?
K
K
K
K
K
K
K

**Figure 3.** Presence of an “embedded code” within the substrate protein sequence. Multiple Lysines may be present in the primary protein sequence. However, typically a one or more select Lysine residues are selected for ubiquitination.

the N-end rule pathway in prokaryotes and eukaryotes employ distinct proteolytic machineries that share common principles of substrate recognition (Mogk et al., 2007). The processes that control N-end have just begun to be unraveled and only a few *in vivo* substrates been identified.

### *PEST sequences*

Particular amino acid sequences within the polypeptide act as proteolytic recognition signals. Analysis of sequence motifs in rapidly degraded proteins, lead Roberts and Rechsteiner to identify PEST sequences. Stretches of PEST sequences which are rich in proline (P), glutamate (E), serine (S), and threonine (T) (along with a lesser extent, aspartic acid) serve as a destruction signal (so called "PEST sequences") (Rogers et al., 1986). Ubiquitination of proteins by multi- subunit ligases, consisting of Ubc3/Cdc34, Skp1, cullin/Cdc53 and F-box proteins, has been shown to be preceded by phosphorylation within the PEST motif (Feldmann et al., 1997). Furthermore, phosphorylation of Ser or Thr residues in the PEST regions of proteins has been shown to activate their recognition and processing by the ubiquitin-proteasome pathway (Yaglom et al., 1995; Lanker et al., 1996; Willems et al., 1996; Won and Reed, 1996**).**

### *D- box and the KEN box*

By far, short sequence motifs serve as primarily signals for degradation. This specific degradation mechanism is involved in regulating cell cycle proteins. Ubiquitination of mitotic cyclins is mediated by a small NH2-terminal motif known as the "destruction box" or “D-box” (Glotzer et al., 1991). The minimal motif is nine residues long with, the following consensus sequence: R-A/T-A-L-G-X-I/V-G/T-N. The destruction box, while either phosphorylated or ubiquitinated serves as a binding site for the ligase subunit of the APC/cyclosome complex. Deletion experiments suggested that $NH_2$-terminal sequences of cyclin B, 90 in sea urchins (Murray and Kirschner, 1989) and 72 in humans (Lorca et al., 1992), play a critical role in targeting cyclins for degradation. The resistance of truncated proteins to degradation indicated interaction of the $NH_2$-terminal portion of cyclin with the destruction machinery. Mutations in the D-box of cyclins severely reduce and/or abolish their ubiquitination abililty (Glotzer et al., 1991; Lorca et al., 1992; Amon et al., 1994; Stewart et al., 1994). Moreover, the cyclin B destruction box is portable, as chimeras containing the N-terminus of cyclin B that has been integrated into other proteins result in their rapid degradation.

A new targeting signal, the KEN box, present in Cdc20 was identified by Pfleger and Kirschner (2000). Mutations studies identified four key residues necessary for substrate recognition in the motif K-E-N-X-X-X-N, (in which aspartic acid in the final position supported similar polyubiquitination as the asparagine). Active KEN boxes have been reported within other proteins and like D-boxes are transposable to other proteins. Both D-box and KEN-box are recognized by Cdh1 and/or Cdc20, which subsequently recruit the APC/cyclosome complex, leading them to ubiquitination and proteasome-mediated degradation of the target protein. The D-box is recognized by both Cdc20 and Cdh1, whereas the KEN-box is preferentially recognized by Cdh1. Cdc20 itself contains a KEN box, which is therefore recognized by Cdh1, ensuring the temporal degradation of Cdc20.

### *Sugar recognition*

N-glycans were recently found to act as ubiquitination signaling molecules. It was recently demonstrated that Fbx2, component of large SCF-type E3 ubiquitin ligase complex specifically binds N-linked glycoproteins and ubiquitinates them, leading to degradation *via* the endoplasmic reticulum associated protein degradation (ERAD) pathway (Yoshida et al., 2002). Fbx2 recognizes high mannose on its substrates to eliminate glycoproteins in neuronal cells. In yeast, the HRD/DER pathway is the main ubiquitination system known to be involved in the ERAD pathway. More E3 ligases outside the HRD/DER pathways are being recognized that target their substrates employing sugar-recognition (Yoshida, 2003).

### *Hydroxyproline*

Hypoxia inducible factor-1 (HIF1) is a heterodimeric transcription factor, composed of alpha and beta subunits, which responds to changes in cellular oxygen content. In the presence of oxygen, HIF1α is targeted for destruction by the E3 Ub ligase VHL.  Human VHL protein recognizes and binds to the conserved hydroxylated proline 564 in the alpha subunit (Ivan et al., 2001). Prolyl hydroxylation of HIF1α by HIF prolyl-hydroxylase is the key regulator of the interaction of the enzyme VHL ligase and HIFα  (Jaakkola et al., 2001). HIF1 is known to play key role in various cellular responses to hypoxia, like the regulation of genes involved in energy metabolism, angiogenesis, and apoptosis. Thus, an absolute requirement for dioxygen as a co-substrate by prolyl-hydroxylase suggests that HIF1 is a master regulator of metabolic adaptation to hypoxia *in vivo* (Semenza, 2000).

### *Protein misfolding*

The molecular chaperones are known to bind misfolded or unfolded proteins to prevent protein aggregation. They either catalyze the refolding of the protein through an ATP-dependent mechanism (if feasible) or target these misfolded proteins for ubiquitination. CHIP (C-terminus of Hsc70-interacting protein) is an excellent example of U-box E3 ligase family as it targets the misfolded proteins (Connell et al., 2001; Jiang et al., 2001). Molecular chaperones such as heat shock protein Hsp70 and Hsp90 work in concert with co-chaperones such as CHIP to promote substrate degradation. CHIP, as mentioned previously, is an E3 ubiquitin ligase enzyme responsible for the ubiquitination of Hsp70 misfolded substrates such as the serine/threonine kinase Raf-1, glucocorticoid receptor, tau and immature CFTR proteins (Connell et al., 2001; Shimura et al., 2004; Petrucelli et al., 2004; Jiang et al., 2001).

### *Phosphorylation based*

Additionally, studies have revealed that a specific ubiquitin ligase recognizes phosphorylated IKBα (pIKBα) through a short peptide stretch, composed of 6 aa motif ( e.g., DS(PO3)GXXS(PO3)). This highly conserved region suggests a well-defined E3 recognition motif. A similar motif is also present in β-catenin, mutating any of the conserved residues within these recognition sites results in stabilization of both IKBs as well as β-catenin. A Lysine residue, located 9–12 aa N-terminal to the recognition site, is also conserved between IKBs and β-catenin, suggesting a single enzyme mediates both the recognition and conjugation of ubiquitin to these substrates *via* two functional sites residing in one or two distinct proteins (Hunter, 2007).

Altogether, these studies illustrate the diversity in determinants of various individual Ub E3 ligases. Thus, there is a need to focus on single Ub E3 ligase system to understand how individual ligases select their targets for modification and achieve site specificity. Numerous large-scale studies have been undertaken to identify ubiquitinated substrates. However, the identification of ubiquitinated Lysines has proven to be difficult for many proteins.

## APPROACHES TAKEN TO IDENTIFY UBIQUITINATED PROTEINS

There is a need for novel techniques designed to identify and characterize protein modifications on a large or global scale. For example, there are more than 500 E3s in the human genome, yet functional information is available for only a small fraction. Linking an E3 with its substrates is difficult and is generally dependent on either a functional connection or a physical association between the proteins. Given the large number of potentially ubiquitinated substrates and E3s, new strategies to deduce E3-substrate pairs are needed since performing biochemical screens for E3 substrates is labor-intensive, is hampered by low substrate levels, as well as, the intrinsically weak interactions between E3s and their substrates.

### *Mass spectrometry approaches*

Most of the studies done to date are either specifically targeted towards identifying the ubiquitinated site in a single protein (like EGFR) or geared toward large-scale approaches ( i.e. identifying the 'ubiquitome' in a cell). These large-scale analyses of ubiquitinated proteins usually employ multi-step approaches that include affinity purification and MS (mass spectrometry) analysis of proteins. This approach was successful in yeast (Peng et al., 2003), human cell lines (Matsumoto et al., 2005), and transgenic mice (Jeon et al., 2007). MS-based approaches to identify precise ubiquitination sites rely on the fact that isopeptide-linked ubiquitin can be cleaved by trypsin between Arg74 and Gly75, producing a signature diglycine peptide.

Ubiquitination can be detected based on two properties; firstly, that peptides containing an ubiquitinated site (or sites) have an incremental molecular mass of 114 Da for each targeted Lysine residue; secondly, that ubiquitin conjugation to a Lysine residue inhibits proteolytic cleavage by trypsin at the modified site. In their landmark approach for large-scale screening of ubiquitinated sites, Peng and colleagues detected 110 ubiquitinated sites from 72 ubiquitin-tagged proteins (Peng et al., 2003). This was the most comprehensive study conducted where endogenous yeast Ub genes were disrupted and replaced by His epitope-tagged ubiquitin. Additionally, their large-scale approach using shotgun sequencing generated a dataset of more than 1000 candidate substrates. Database searching revealed 110 ubiquitinated sites on 72 different proteins. Subsequently, use of tagged ubiquitin *in vivo* in a transgenic mouse model was described (Tsirigotis et al., 2001). Immunoaffinity purification of ubiquitinated substrates in mammals (Vasilescu et al., 2005) was used to separate substrates after being trypsinized. Over 70 ubiquitinated proteins and 16 signature Ub attachment sites were identified by LC-MS/MS analysis. In a variation of this method, identified potential Ub ligase substrates were identified by subjecting the immunoaffinity purified fractions from human cells to both native and denaturing conditions (Matsumoto et al., 2005). Combinations of several proteomic studies are summarized with regard to the purification strategies, methods used and total number of Ub-tagged candidates identified (Table 1).

While recent advances in mass spectrometry have quickly expanded the repository of proteins modified by the ubiquitin family, MS-based approaches are still biased towards identifying highly abundant and stable complexes. Ub ligase-substrate complexes are known to be transient and only a fraction of the sampled protein is ubiquitinated at a given time. Also, it has been reported that miscleavage at Arg74 in the ubiquitin sequence generates a longer tag (LRGG) that is difficult to identify. The peptides generated by trypsin sometimes are too large to undergo standardized analytical procedures. Most of the purification strategies use tagged ubiquitin, but there are still no reports on how ubiquitination machinery reacts towards tagged ubiquitin as compared to the wild-type. Moreover the accurate identification of Ub substrates is hindered because some ubiquitin-like proteins (Nedd8 and ISG15) are known to target Lysine residues which are known to generate the same GG peptides by trypsin digestion, as with ubiquitin. This results in detection of false positive results. Thus, MS-based proteomics identifies a broad range of post-translationally modified substrates in an unbiased manner. In addition to this, only relatively few ubiquitinated substrates have been identified due to the difficulty of detecting small quantities of transient Ub-tagged proteins in the complex mixed with highly abundant proteins in the purified sample. This requires an additional step in the identification procedure in order to separate out those proteins from ubiquitinated samples. While various fractionation studies have been applied prior to MS to overcome these barriers, there still exist issues regarding resolution and sample loss. Thus, despite the extensive efforts to accurately identify Ub substrates and the target site, the MS-based methods used have been laborious and results far from accurate. As a result novel methods like stable-isotope-based quantification strategies and development of non-MS based approaches to aid in differentiating Ub-targeted proteins from the background proteins without the need to enrich ubiquitinated substrate pool in the sample is much needed.

### *Non-mass spectrometry approaches*

Another approach toward developing tools for the purification of ubiquitinated substrates is making use of the fact that UBA domains bind polyubiquitin chains with high affinity. The relative ease of UBA–agarose conjugates production, as compared with anti-ubiquitin antibody production, makes these domains an attractive resource in ubiquitin pull-down experiments.

Ubiquitin-binding proteins have been described based on the type of ubiquitin-binding domains/motifs they possess. Their ubiquitin-binding properties have just begun to be exploited in charactering the ‘ubiquitome’, which consists of all ubiquitinated proteins in the cell. The ability of the UBA domain to bind polyubiquitin was employed in a screen coupled with *in vitro* transcription/translation of a human cDNA library from adult brain to identify proteins interacting with the p62 UBA domain (Pridgeon et al., 2003). A total of 11 proteins were identified as putative ubiquitinated proteins, most of which were important in neuropathologies. With approximately 5% of the total *Arabidopsis* proteins known to be involved in the UPS/proteasome system, more and more studies are being directed towards identifying ubiquitinated substrates. The first large scale study conducted in plants used recombinant GST-tagged ubiquitin binding domains (UIM and double UBA domain). Affinity purified ubiquitinated proteins were separated by SDS-PAGE, and then trypsin-digested before they were analyzed by a multidimensional protein identification technology (MudPIT) system; more than 290 putative ubiquitinated proteins were identified and 85 ubiquitinated Lysine residues in 56 proteins were characterized (Maor et al., 2007). More recently, affinity purification employing the UBA domain of p62 yielded a total of 200 putative ubiquitinated proteins from *Arabidopsis* (Manzano et al., 2008). Proteins bound to the p62-agarose matrix were digested with trypsin and later separated by HPLC chromatography followed by identification by MALDI-TOF/TOF. However, affinity purification of ubiquitinated substrates, using a UBA domain has its drawbacks. Apart from interacting with ubiquitin, some UBA domains interact with UBL domains (Walters et al., 2003; Lowe et al., 2006; Kang et al., 2007; Layfield et al., 2001), as well as, other proteins (Dieckmann et al., 1998; Feng et al., 2004; Gao et al., 2003; Boutet et al., 2007; Gwizdek et al., 2006; Ota et al., 2008), thus raising questions regarding their specificity with respect to ubiquitin chains. A combination of SILAC (stable isotope labeling with amino acids in cell culture), parallel affinity purification (PAP), and mass spectrometry was used to identify F-box ligase substrates in yeast. This approach was successful in identifying transiently modified substrates and proteins tagged with poly Lys-48 chains for degradation; however, this method failed to detect already reported substrates such as Fzo1p (Fritz et al., 2003; Escobar-Henriques et al., 2006; Cohen et al., 2008), and Gal4p (Muratani et al., 2005).

Using a yeast protein microarray numerous known and novel ubiquitinated substrates of the E3 ligase Rsp5 were recently identified in a high-throughput manner (Gupta et al., 2007). These protein microarrays contained more than 4000 GST- and 6 × HIS-tagged yeast proteins from *S. cerevisiae* spotted on nitrocellulose slides and directly tested for ubiquitination by Rsp5 *in vitro*. However, not all known Rsp5 substrates were identified in their screen, since some of the known substrates were not printed on the array, and some Rps5 substrates are known to require adaptor proteins to bind to Rsp5. Moreover, there is a possibility that some of the substrates might have been lost in the purification process because of their weak and transient interaction with the enzyme, making it impossible to determine the impact the tags had on the accessibility of some substrates. A more powerful approach, global protein stability (GPS) profiling consists of a fluorescence-based multiplex system for assessing protein stability on a high-throughput scale for SCF substrates (Yen and Elledge, 2008). A powerful feature of this technique was that it monitored the E3 ligase activity. This screen recovered 73% of the previously reported SCF substrates and found a total of 359 proteins as likely substrates.

**Table 1.** Comparison of Mass-spectrometric approaches and non-spectrometric approaches todentify ubiquitinated proteins and target sites.

| **Mass spectrometric approaches** | | | |
|---|---|---|---|
| ***Purification strategies*** | ***Screen*** | ***Substrates/sites identified*** | ***References*** |
| (HIS)$_6$-biotin-Ub<br>Ni-chelate chromatography<br>LC/LC-MS/MS | Hela cells | 100 proteins<br>Included both ubiquitinated ubiquitin associated proteins | Gururaja et al. |
| Membrane associated | Yeast proteome | 211 overall identified<br>83 prtoeins ERAD substrates<br>> 30 sites | Hitchcock et al. |
| FT-ICR MS | Ubc5 | 15 sites | Cooper et al. |
| In gel digestion<br>LC-MS/MS | Breast cancer cells | 96 sites | Denis et al. |
| SCX cation exchange<br>LC/LC-MS/MS | Yeast proteome | 1075 proteins<br>110 sites | Peng et al. |
| No Ub tag<br>Immunoaffinity<br>GeLC-MS/MS | Breast cancer cells | 70 proteins | Vasilescu et al. |
| No Ub tag<br>Immunoaffinity with (native and denaturing)<br>LC/LC-MS/MS | Human cells | proteins identified<br>670 native conditions<br>345- denaturing conditions<br>18 sites | Matsumoto et al. |
| MALDI-TOF MS/MS of sulfonated tryptic peptides | CHIP | 3 proteins<br>1 site | Wang et al. |
| In vitro Ub assay | BRAC1/BARD1 | 2 proteins | Sato et al.<br>Starita et al. |
| (HIS)$_6$-biotin-Ub<br>Native nickel chromatography<br>LC/LC-MS/MS | Human cells | 22 proteins<br>4 sites | Kirkpatrick et al. |
| Subtractive Ub profiling<br>Affinity purification<br>LC/LC-MS/MS | Proteasome receptor Rpn10 in Yeast | 54 substrates | Mayor et al. |

| **Non- Mass spectrometric approaches** | | | |
|---|---|---|---|
| ***Purification strategies*** | ***Screen*** | ***Substrates/sites identified*** | ***References*** |
| Two-hybrid screen | Yeast proteome | Some positive substrates | Uetz et al. |
| Luminescent assay<br>Ub-biotin | 188 purified GST-tagged yeast proteins | 7 novel Rsp5 substrates | Kus et al. |
| Protein Microarrays | Yeast proteome | 150 potential substrates<br>40 strong candidates | Gupta et al. |
| UBA-association | Adult human brain cDNA library screen | 11 proteins | Pridgeon et al. |
| S5a-affinity chromatography<br>Two-dimensional analysis | Mammalian tissues | Some proteins<br>hHR23B identified | Layfield et al. |
| Affinity purification<br>GST-fused UBDs<br>LC-MS/MS-based (MudPIT) analysis | *Arabidopsis* proteome | 294 proteins<br>85 sites | Moar et al. |

Since the technique measured indirect effects of the SCF ligase activity on proteins, all those proteins whose stability was either increased or decreased in response to various drugs or stimuli were reported. However, the GPS technique can failed to detect a protein whose functionality was altered as a result of ubiquitination, or if a protein changed its localization in the cell or acquired different binding partners. Again, it was impossible to access what role the fusion tag may have played in the stability of these proteins.

Recent advances in this field have been made by the generation of antibodies that are capable of recognizing ubiquitin linkages of a specific conformation. Two groups have independently generated K63-chain specific antibodies for use in Western blotting (Newton et al., 2008; Wang et al., 2008). These reagents should enhance the identification of K63 ubiquitinated substrates and further define the functional role for this tag.

Clearly, it has been difficult to achieve a robust approach for the large-scale identification of ubiquitinated substrates in the cell. Each of the methods employed to date have inherent advantages and disadvantages, therefore there is a need for an alternative solution toward solving the problem of identifying the "embedded code" that predicts Lysine selectivity in a target substrate.  Lessons can be learnt from computational investigations aimed at identification of a SUMOylation motif required for target selection (Rodriguez et al., 2001).

## LESSONS FROM SUMO: EXAMINING THE NEAREST KIN

Of the several new Ubl modifiers that have been discovered in the past few years, the SUMO pathway has received the most intense scrutiny. SUMO was identified in 1996 as a peptide conjugated to the nucleocytoplasmic-transport protein RanGAP1, resulting in a change in its cellular localization (Matunis et al., 1996). Since the discovery of SUMO as a post-translational protein modifier over 10 years ago, more than 200 proteins targets have been reported, with the majority being nuclear proteins. SUMOylation is known to cause either alteration in protein localization, a change in protein activity, or differences in interaction with binding partners (Geiss-Friedlander and Melchior, 2007). SUMO is about 20% similar to ubiquitin in its primary sequence and contains ~15 additional N-terminal amino acid residues (Bayer et al., 1998). Like, ubiquitination, SUMOylation is achieved by sequential action of three enzymes; the activating (E1), conjugating (E2), and ligating (E3) enzymes. Nevertheless, SUMO E1, E2, and E3s are very distinct from the E1, E2 and E3 of the ubiquitination system (Yeh et al., 2000). Despite the similarities in structure and conjugation mechanism, they both have distinct physiological effects in the cell. To date, there is only one reported example of both E1 (SAE1/SAE2 heterodimer) and E2 (UBC9) for SUMOylation, in contrast to the large number of E1s and E2s reported for the ubiquitination pathway. Like the ubiquitination system several SUMO E3 ligases have been identified, most of which have a SiYz/PIAS (SP)-ring motif required for their function. There are three types of known SUMO E3 ligases – PIAS proteins, RanBP2, and Pc2 each conferring substrate specificity to the SUMOylation reaction.

As additional SUMO targets and pathways influenced by SUMO regulation are recognized, the significance of this pathway is beginning to be appreciated. SUMOylation is known to participate in diverse cellular events, including chromosome segregation and cell division, DNA replication and repair, transcriptional regulation, nuclear transport and signal transduction (Müller et al., 2001). Four different type of SUMO isoforms (SUMO1 - 4) are reported in mammals. SUMO-1 is the most commonly found conjugated isoform under normal conditions. SUMO-2 and SUMO-3 have very similar sequence identity and appear to be conjugated in response to stress signals. SUMO-4 is more tissue-specific, as it is identified in human kidney, suggesting its involvement in more tissue-dependent functions. Both SUMO2/3 and SUMO-4 contain an internal consensus motif ΨKXE (where Ψ represents a large hydrophobic amino acid, and X represents any amino acid) that is required for SUMO modification both *in vivo* and *in vitro* (Rodriguez et al., 2001), which is missing in SUMO-1. Exploiting the fact that Ubc9 binds to this motif directly (Sampson et al., 2001), a number of SUMO targets have been identified *via* their interaction with Ubc9 in the yeast two-hybrid screen. Not all ΨKXE motif found in proteins are modified, as SUMO E3s are presumed to enhance specificity by interacting with other features of the substrate. In addition, to the consensus sequence amino acids upstream or downstream of the acceptor Lysine may help to insure accessibility of the substrate for the conjugation apparatus. For some SUMO substrates, additional interactions occur outside the consensus sequence (Anckar and Sistonen, 2007; Bernier-Villamor et al., 2002), demonstrating the involvement of multiple, co-operating interactions in regulating the target selection process. In this regard, the consensus sequence can be seen as a local mediator of substrate-conjugation apparatus interaction, fine-tuning the SUMO conjugation event by facilitating the correct positioning of the target Lysine residue to the active site of Ubc9.

Approaches similar to the identification of ubiquitinated substrates have been utilized in identifying novel SUMO targets and/or total SUMOylated substrates in the cell. These methods rely upon purification of SUMOylated proteins from cell lysates *via* affinity tags, followed by MS analysis (Li et al., 2004; Zhao et al., 2004; Zhou et al., 2004; Vertegaal et al., 2004; Wohlschlegel et al., 2004; Panse et al., 2004). A variety of affinity-tagged SUMOs have been described that have been overexpressed to overcome low levels of SUMOylated proteins in the cells, a major barrier to MS sensitivity. Moreover, at a given time only a small fraction of proteins in the cells are SUMOylated, since it is a dynamic process in which conjugation and de-conjugation work in concert. It has been suggested that <1% of the proteins in a cell are SUMO modified at any given time (Johnson, 2004), thus making efforts at detecting these modified proteins difficult. The use of several genomic/proteomic and *in silico* combinatorial approaches to identify global pool of ‘Sumo-tome’ has lead to identification of ~500 potential SUMO substrates (Wohlschlegel at el., 2004; Gocke et al., 2005; Zhou et al., 2005). However, *bona fide* SUMOylation sites may still remain to be identified or confirmed *in vivo*. Thus, as experimental proteomics approaches become more and more-labor intensive and time-consuming, there is a growing need to develop prediction tools that would aid in successfully predicting the target substrate.

In this regard, computational techniques have presented a promising approach toward identifying SUMOylation sites. Given this, the first computational prediction tool SUMOplot, was developed which predicted the probability for a SUMO attachment. The SUMOplot prediction heavily depended on identification of the SUMO consensus motif. This limited the prediction results as many non-consensus true positives were missed. SUMOsp was developed based on a manually curated 239 experiment-verified SUMOylation sites from the literature (Xue et al., 2006). GPS and MotifX, two earlier described strategies, were applied to the dataset, yielding good (89.12%) prediction platform for SUMOylation sites. Another bioinformatic study to accurately predicted SUMO modified sites employing a statistical method based on properties of individual amino acid surrounding the SUMO site (Xu et al., 2008).

## STATUS QUO ON UBIQUITINATION SITES

To better understand Lysine selectivity within a protein destined for ubiquitination (Figure 3), it is first important to survey the literature for reported proteins and their ubiquitination sites. The first report exploring the preferences for a specific ubiquitination site was conducted on human red blood cell protein α-spectrin (Galluzzi et al., 2001). The investigators demonstrated that the leucine zipper was a potential ubiquitin recognition motif by site-directed mutagenesis. Moreover, in addition to the primary sequence it has been suggested that secondary folding also plays a role in directing the Lysine selected for ubiquitination. The leucine zipper described in multi-ubiquitination of c-Jun (Treir et al., 1994) is observed in a number of other gene regulatory proteins with 75% similarity to the flanking regions of ubiquitinated α-spectrin Lysine (Murantani and Tansey, 2003). This suggests a conformational recognition mechanism in which positioning of the Lys plays an important role in directing specificity. In another study, K187 (out of the possible six available Lysines) was found to be a preferred ubiquitin target site in the transcription activator Rpn4 (Ju and Xie, 2006). Primary sequence analysis revealed the close proximity of K187 to the N-terminal acidic domain, which acts as ubiquitination signal for transcription activators. Additionally, surface hydrophobic residues are known to be required for ubiquitination of several proteins for proteasomal degradation (Bogusz et al., 2006; Johnson et al., 1998). The neurotrophin receptor TrkA was one of the first receptors to be identified as a K63-polyubiquitin tagged at K485 (Geetha et al., 2005). Recently, ubiquitination of a Lysine within the membrane proximal region of granulocyte colony-stimulating factor receptor (G-CSFR) was reported (Wolfler et al., 2009) and K63-ubiquitination of K338 was reported for the Jen1 Transporter (Paiva et al., 2009) Altogether, a picture is emerging where K63-chains may play a role in regulating internalization and sorting of receptors.

Studies conducted on both the Huntingtin and Androgen receptors support the importance of conserved pentapeptide pattern (FQXL(L/F)) as determinants in their degradation by the proteasome (Chandra et al., 2008). Another report on the E3 substrate selection process analyzed the ubiquitinized-yeast proteome based on subcellular localization (Catic et al., 2004). This study revealed the presence of compartment-specific sequence patterns for ubiquitinated substrates. Structural analyses of ubiquitinated proteins demonstrate a preference for an exposed Lysine residue on the surface of the molecule. Additionally, a survey of 40 ubiquitination sites from 23 proteins showed clear secondary structure preference for Lysine ubiquitination. Modifications were prominent at the Lysines occurring in loop regions (26/40) followed by Lysines in α-helices (10/40) (Catic et al., 2004). This investigation also reported the presence of compartment-specific motifs within the dataset. For example, nuclear proteins had preference for ubiquitination of Lysines near the phosphorylatable residues. Similar bias was observed for ubiquitinated plasma membrane proteins that had either Glu or Asp at -1 or -2 positions from the acceptor Lysine (Catic et al., 2004). Thus, investigating the overall primary and secondary structure as well as the proteins' subcellular localization could yield important information regarding the targeting of the substrates.

## SPECIFICITY PROVIDED BY A SCAFFOLD

Many E3 ligases are known to interact with specific substrates either directly or through scaffold proteins. Scaffold proteins facilitate interaction between the E3 enzymes and their substrates through their multi-domain architecture. One such scaffold is p62, a highly conserved and transcriptionally regulated protein that plays important roles in ubiquitination, receptor trafficking, protein aggregation, and inclusion formation (Seibenhener et al., 2004). P62 acts as a scaffold by interacting with the RING E3, TRAF6, through a TRAF-binding site (TBS) as well as other proteins through one of its many protein-protein interaction domains. Interaction between p62 and TRAF6 has been shown to auto-activate TRAF6 (Wooten et al., 2001; 2006). Functional domains in p62 include a Phox and Bem1p (PB1) domain, a TRAF6-binding region, and an UBA domain (Geetha et al., 2002). The C-terminal UBA domain of p62 has been shown to non-covalently bind ubiquitin (Mueller et al., 2002). Moreover, p62 functions as a shuttling factor for polyubiquitinated substrates by binding the ubiquitinated proteins through its UBA domain and the 26S proteasome through its N-terminal PB1 domain (Wooten et al., 2005). The tyrosine kinase receptor A (TrkA) (Geetha et al., 2005) and the neurotrophin receptor interacting factor (NRIF) (Geetha et al., 2005), both have been shown to be K63- polyubiquitinated by the TRAF6/p62 complex.  In a recent study, in a attempt to understand the Lysine selection process employed by TRAF6/p62 the primary sequences of the Lysines that were targeted for ubiquitination in both TrkA and NRIF were examined for a possible consensus motif (Jadhav et al., 2008). A close look at these two substrates revealed the presence of a conserved consensus pattern for ubiquitination by the TRAF6/p62 complex. This consensus pattern has also been observed in others members of the Trk receptor family, TrkB and TrkC (Jadhav et al., 2008). Interestingly a consensus pattern identified in these proteins was a 10-amino acid long stretch {[– (hydrophobic) – k – (hydrophobic) – x – x – (hydrophobic) – (polar) – (hydrophobic) – (polar) – (hydrophobic)] where k was the ubiquitinated Lysine residue and x any other amino acid} required to successfully target the primary Lysine residue (Jadhav et al., 2008). These studies further suggest the possibility that an "embedded code" that exists whereby an E3 ligase targets a specific Lysine residues for modification over others. Therefore, to better understand the Lysine selection process during ubiquitination, it is important to examine the enzyme-specific selection process. The development of an algorithm to search a training dataset of p62/TRAF6 interactors could be employed as a first step in development of a computational tool to aid in discovery of TRAF6 targets.

## MODEL FOR SUBSTRATE SELECTION

Substrate selection and site specificity is a multi-step process depending on two types of signals, both primary and secondary. The primary signals are the structural motifs; α-helices or β-sheets that influence the local architecture of the primary sequence. Secondary signals, on the other hand, are inherent primary sequences that are essential for the recognition of the primary ubiquitination site. Of both, secondary signals can vary slightly depending on the localization of proteins in the cell.

What can be learned from the E3 TRAF6? In the case of TrkA site-specific ubiquitination (Geetha et al., 2005), the E3, TRAF6, exists as a complex with the E2, UbcH7, in the cytosol. Post-receptor stimulation, the E2/E3 pair form a transient complex recruited to the scaffold, p62, to mediate the ubiquitination of TrkA (Geetha et al., 2005). The target Lysine within a protein can either be buried inside a hydrophobic pocket of the globular protein structure or masked, while the protein is interacting with a different binding partner. Binding of the scaffold protein likely induces a conformational change in the proteins' structure exposing the buried target site (Figure 4A). Thereafter, the scaffold recruits the activated E3/E2 complexes to the substrate protein. The enzyme complex then scans the exposed surface for an acceptor Lysine that possesses the appropriate conformation. Once an accessible Lysine is recognized and if the nearby flanking residues present an appropriate environment, transfer of the ubiquitin molecule occurs. In other cases, the active enzyme complex E3/E2 first binds to the substrate protein and produces a similar type of conformational change (i.e., exposure of the target site). This binding of substrate to the E3 produces structural changes for accommodating the scaffold protein to the complex, which aids in the enzymatic process (Figure 4B). These results suggest that the former model is more likely operative for site-specific ubiquitination of the target (Geetha et al., 2005).

## SUMMARY

The analysis of the ‘ubiquitome’ presents one of the most exciting and challenging tasks in current proteomics research. The ultimate limiting factor in studying ubiquitination substrate selection mechanism is the lack of curated data sets of ubiquitinated proteins. This makes it difficult to evaluate, and compare target sites to decode selectivity and specificity. With identification of more than 500 or so ubiquitin ligases there exists a need to rapidly and precisely identify enzyme-specific substrates. This task demands that we take multiple novel approaches as well as a combination of techniques to precisely identify target sites for these ligases. With rapid advancement in mass spectrometric analysis and more sophistication in proteomic tools and novel approaches we can expect the number of precisely identified sites to rise. Moreover, use of bioinformatic methods to predict site modification *in silico* could yield more efficient results. These prediction tools should be closely integrated into the interpretation of proteomic experiments. Also as proteomics methods identify more and more *in vivo* ubiquitination sites, prediction algorithms can be fine tuned and improved with this information. The model that I propose here can be applied to other E3 Ub ligases that are known to employ scaffold proteins to aid in their substrate selection process (Figure 4). For example, the BTB-domain proteins that were identified as substrate-specific scaffolds for Ub E3 ligase CUL-3 *in C. elegans* (Xu et al.,

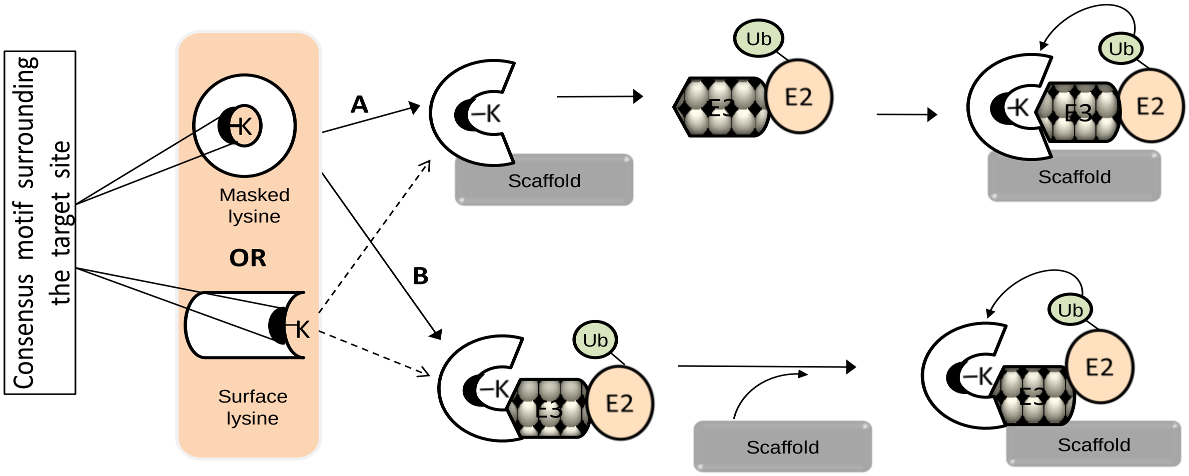

Consensus motif surrounding the target site
K
Masked lysine
OR
K
Surface lysine
A
B
K
Scaffold
Ub
E3
E2
Ub
K
E3
E2
Scaffold
Ub
K
E3
E2
Scaffold
Ub
K
E3
E2
Scaffold

**Figure 4.** Model for substrate selection mechanism for Ub E3 ligase/scaffold complex. The target Lysine site can either be masked or buried inside the hydrophobic pocket of the globular protein structure or be exposed to the exterior surface on the substrate. A) The scaffold protein interacts with the E3/E2 complex providing specificity for ubiquitination. Employing an embedded code the complex, with the assistance of the scaffold, directs ubiquitination of the target substrate on one or more specific Lysine residues. This model is supported by studies with p62/TRAF6 complex (Geetha et al., 2005).  B) Alternatively, the interaction of the E3 with the putative substrate changes the conformation of the substrate and allows it to recruit scaffold protein which in turn provides a platform for the ubiquitination reaction to take place.

2003). Lysine ubiquitination interplays actively with other post-translational modifications, either agonistically or antagonistically, to form a coded message for intramolecular signaling programs that are crucial for governing cellular functions. Given the intricacy of the ubiquitin system, research into its functions and mechanisms should continue to yield novel insights into cell regulation.

## ACKNOWLEDGEMENTS

This work was funded by NIH (NINDS 33661) to MWW. We thank Drs. Scott Santos and Michael Wooten for reading and review of draft form of this manuscript.

# REFERENCES

1. Adhikari A, Chen ZJ (2009) Diversity of polyubiquitin chains. Developmental Cell 16:485-486.

2. Amon A, Irniger S, Nasmyth K (1994) Closing the cell cycle circle in yeast: G2 cyclin proteolysis initiated at mitosis persists until the activation of G1 cyclins in the next cycle. Cell 77: 1037–1050.

3. Anckar J, Sistonen L (2007) SUMO: getting it on. Biochem Soc Trans 35: 1409-1413.

4. Aravind L, Koonin EV (2000) The U box is a modified RING finger - a common domain in ubiquitination. Curr Biol 10: R132-R134.

5. Arnason T, Ellison MJ (1994) Stress resistance in *Saccharomyces cerevisiae* is strongly correlated with assembly of a novel type of multiubiquitin chain. Mol Cell Biol 14: 7876–7883.

6. Aviel S, Winberg G, Massucci M, Ciechanover A (2000) Degradation of the Epstein-Barr virus latent membrane protein 1 (LMP1) by the ubiquitin-proteasome pathway. Targeting *via* ubiquitination of the N-terminal residue. J Biol Chem 275: 23491–23499.

7. Bachmair A, Finley D, Varshavsky A (1986) In vivo half-life of a protein is a function of its amino-terminal residue. Science 234: 179−186.

8. Bayer P, Arndt A, Metzger S, Mahajan R, Melchior F, Jaenicke R, Becker J, (1998) Structure determination of the small ubiquitin-related modifier SUMO-1. J Mol Biol 280: 275–286.

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

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

11. Bloom J, Amador V, Bartolini F, DeMartino G, Pagano M (2003) Proteasome-mediated degradation of p21 *via* N-terminal ubiquitinylation. Cell 115: 71-82.

12. Bogusz M, Brickley DR, Pew T, Conzen SD (2006) A novel N-terminal hydrophobic motif mediates constitutive degradation of serum- and glucocorticoid-induced kinase-1 by the ubiquitin-proteasome pathway. FEBS J 273: 2913-2928.

13. Boutet SC, Disatnik MH, Chan LS, Iori K, Rando TA (2007) Regulation of Pax3 by proteasomal degradation of monoubiquitinated protein in skeletal muscle progenitors. Cell 130: 349–362.

14. Breitschopf K, Bengal E, Ziv T, Admon A, Ciechanover A (1998) A novel site for ubiquitination: the N-terminal residue, and not internal Lysines of MyoD, is essential for conjugation and degradation of the protein. EMBO J 17: 5964–5973.

15. Cadwell K, Coscoy L (2005) Ubiquitination on nonLysine residues by a viral E3 ubiquitin ligase. Science 309: 127-130.

16. Catic, Collins C, Church GM, Ploegh HL (2004) Preferred *in vivo* ubiquitination sites. Bioinformatics 20: 3302-3307.

17. Chandra S, Shao J, Li JX, Li M, Longo FM, Diamond MI (2008) A common motif targets Huntingtin and the Androgen receptor to the proteasome. J Biol Chem 283: 23950-23955.

18. Chau V, Tobias JW, Bachmair A, Marriott D, Ecker DJ, Gonda DK, Varshavsky A (1989) A multiubiquitin chain is confined to specific Lysine in a targeted short-lived protein. Science 243: 1576–1583.

19. Cohen MM, Leboucher GP, Livnat-Levanon N, Glickman MH, Weissman AM (2008) Ubiquitin-proteasome-dependent degradation of a mitofusin, a critical regulator of mitochondrial fusion. Mol Biol Cell 19: 2457–2464.

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

21. Cooper HJ, Heath JK, Jaffray E, Hay RT, Lam TT, Marshall AG (2004) et al. Identification of sites of ubiquitination in proteins: a fourier transform ion cyclotron resonance mass spectrometry approach. Anal Chem 76: 6982−6988.

22. Dieckmann T, Withers-Ward ES, Jarosinski MA, Liu CF, Chen IS, Feigon J (1998) Structure of a human DNA repair protein UBA domain that interacts with HIV-1 Vpr. Nat Struct Biol 5: 1042–1047.

23. Escobar-Henriques M, Westermann B, Langer T (2006) Regulation of mitochondrial fusion by the F-box protein Mdm30 involves proteasome-independent turnover of Fzo1. J Cell Biol 173: 645–650.

24. Feldmann RMR, Correll CC, Kaplan KB, Deshaies RJ (1997) A complex of Cdc4p, Skp1p and Cdc53p/Cullin catalyzes ubiquitination of the phosphorylated inhibitor Sic1p. Cell 91: 221-230.

25. Feng P, Scott CW, Cho NH, Nakamura H, Chung YH, Monteiro MJ, Jung JU (2004) Kaposi's sarcoma-associated herpesvirus K7 protein targets a ubiquitin-like/ubiquitin-associated domain-containing protein to promote protein degradation. Mol Cell Biol 24: 3938–3948.

26. Fritz S, Weinbach N, Westermann B (2003) Mdm30 is an F-box protein required for maintenance of fusion-competent mitochondria in yeast. Mol Biol Cell 14: 2303–2313.

27. Galluzzi L, Paiardini M, Lecomte MC, Magnani M (2001) Identification of the main ubiquitination site in human erythroid alpha-spectrin. FEBS Lett 489: 254-258.

28. Gao L, Tu H, Shi ST, Lee KJ, Asanaka M, Hwang SB, Lai MM (2003) Interaction with a ubiquitin-like protein enhances the ubiquitination and degradation of hepatitis C virus RNA-dependent RNA polymerase. J Virol 77: 4149–4159.

29. Geetha T, Jiang J, Wooten MW (2005) Lysine 63 polyubiquitination of the nerve growth factor receptor TrkA directs internalization and signaling. Mol Cell 20: 301–312.

30. Geetha T, Kenchappa RS, Wooten MW, Carter BD (2005) TRAF6-mediated ubiquitination regulates nuclear translocation of NRIF, the p75 receptor interactor. EMBO J 24: 3859–3868.

31. Geetha T, Wooten MW (2002) Structure and functional properties of the ubiquitin binding protein p62. FEBS Lett 512: 19-24.

32. Geiss-Friedlander R, Melchior F (2007) Concepts in sumoylation: a decade on. Nat Rev Mol Cell Biol 8: 947–956.

33. Glotzer M, Murray AW, Kirschner MW (1991) Cyclin is degraded by the ubiquitin pathway. Nature 349: 132-138.

34. Gocke CB, Yu H, Kang H (2005) Systematic identification and analysis of mammalian small ubiquitin-like modifier substrates. J Biol Chem 280: 5004–5012.

35. Gupta R, Kus B, Fladd C, Wasmuth J, Tonikian R, Sidhu S, Krogan NJ, Parkinson J, Rotin D (2007) Ubiquitination screen using protein microarrays for comprehensive identification of Rsp5 substrates in yeast. Mol Syst Biol 3: 116.

36. Gururaja T, Li W, Noble WS, Payan DG, Anderson DC (2003) Multiple functional categories of proteins identified in an *in vitro* cellular ubiquitin affinity extract using shotgun peptide sequencing. J Proteome Res 2: 394-404.

37. Gwizdek C, Iglesias N, Rodriguez MS, Ossareh-Nazari B, Hobeika M, Divita G, Stutz F, Dargemont C (2006) Ubiquitin-associated domain of Mex67 synchronizes recruitment of

the mRNA export machinery with transcription. Proc Natl Acad Sci USA 103: 16376–16381.

38. Hershko A, Ciechanover A (1998) The ubiquitin system. Annu Rev Biochem 67: 425-479.

39. Hicke L (1999) Gettin' down with ubiquitin: turning off cell-surface receptors, transporters and channels. Trends Cell Biol 9: 107–112.

40. Hitchcock AL, Auld K, Gygi SP, Silver PA (2003) A subset of membrane-associated proteins is ubiquitinated in response to mutations in the endoplasmic reticulum degradation machinery. Proc Natl Acad Sci USA 100: 12735-12740.

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

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

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

44. Hunter T (2007) The age of crosstalk: phosphorylation, ubiquitination, and beyond. Mol Cell 28: 730-738.

45. Hurley JH, Lee S, Prag G (2006) Ubiquitin-binding domains. Biochem J 399: 361-372.

46. Ivan M, Kondo K, Yang H, Kim W, Valiando J, Ohh M, Salic A, Asara JM, Lane WS, Kaelin WG Jr. (2001) HIF-alpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O(2) sensing. Science 292: 464-468.

47. Jaakkola P, Mole DR, Tian YM, Wilson MI, Gielbert J, Gaskell SJ, Kriegsheim AV, Hebestreit HF, Mukherji M, Schofield CJ, Maxwell PH, Pugh CW, Ratcliffe PJ (2001) Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. Science 292: 468-472.

48. Jadhav T, Geetha T, Jiang J, Wooten MW (2008) Identification of a consensus site for TRAF6/p62 polyubiquitination. Biochem Biophys Res Commun 371: 521-524.

49. Jeon HB, Choi ES, Yoon JH, Hwang JH, Chang JW, Lee EK, Choi HW, Park Z, Yoo YJ (2007) A proteomics approach to identify the ubiquitinated proteins in mouse heart. Biochem Biophys Res Commun 357: 731–736.

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

51. Johnson ES (2004) Protein modification by SUMO. Annu Rev Biochem 73: 355-382.

52. Johnson PR, Swanson R, Rakhilina L, Hochstrasser M (1998) Degradation signal masking by heterodimerization of MATalpha2 and MATa1 blocks their mutual destruction by the ubiquitin-proteasome pathway. Cell 94: 217-227.

53. Ju D, Xie X (2006) Identification of the preferential ubiquitination site and ubiquitin-dependent degradation signal of Rpn4. J Biol Chem 281: 10657-10662.

54. Kaiser P, Huang L (2005) Global approaches to understanding ubiquitination. Genome Biol 6: 233-211.

55. Kang Y, Zhang N, Koepp DM, Walters KJ (2007) Ubiquitin receptor proteins hHR23a and hPLIC2 interact. J Mol Biol 365: 1093–1101.

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

57. Kuhlbrodt K, Mouysset J, Hoppe T (2005) Orchestra for assembly and fate of polyubiquitin chains. Essays Biochem 41: 1-14.

58. Kus B, Gajadhar A, Stanger K, Cho R, Sun W, Rouleau N, Lee T, Chan D, Wolting C, Edwards A, Bosse R, Rotin D (2005) A high throughput screen to identify substrates for the ubiquitin ligase Rsp5. J Biol Chem 280: 29470-29478.

59. Lanker S, Valdivieso MH, Wittenberg C (1996) Rapid degradation of the G1 cyclin Cln2 induced by CDK-dependent phosphorylation. Science 271: 1597-1600.

60. Layfield R, Tooth D, Landon M, Dawson S, Mayer J, Alban A (2001) Purification of poly-ubiquitinated proteins by S5a-affinity chromatography. Proteomics 1: 773–777.

61. Li T, Evdokimov E, Shen R-F, Chao C-C, Tekle E, Wang T, Stadtman ER, Yang DCH, Chock PB (2004) Sumoylation of heterogeneous nuclear ribonucleoproteins, zinc finger proteins, and nuclear pore complex proteins: A proteomic analysis. Proc Nat. Acad Sci USA 101: 8551–8556.

62. Li W, Bengtson MH, Ulbrich A, Matsuda A, Reddy VA, Orth A, Chanda SK, Batalov S, Joazeiro CA (2008) Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling. PLoS ONE 3: e1487.

63. Lorca T, Devault A, Colas P, Van Loon A, Fesquet D, Lazaro JB, Dorée M (1992) Cyclin A-Cys41 does not undergo cell cycle-dependent degradation in Xenopus extracts. FEBS Lett 306: 90–93.

64. Lowe ED, Hasan N, Trempe JF, Fonso L, Noble ME, Endicott JA, Johnson LN, Brown NR (2006) Structures of the Dsk2 UBL and UBA domains and their complex. Acta Crystallogr D Biol Crystallogr 62: 177–188.

65. Manzano C, Abraham Z, López-Torrejón G, Del Pozo JC (2008) Identification of ubiquitinated proteins in *Arabidopsis*. Plant Mol Biol 68: 145-158.

66. Maor R, Jones A, Nühse TS, Studholme DJ, Peck SC, Shirasu K (2007) Multidimensional protein identification technology (MudPIT) analysis of ubiquitinated proteins in plants. Mol Cell Proteomics 6: 601-610.

67. Matsumoto M, Hatakeyama S, Oyamada K, Oda Y, Nishimura T, Nakayama KI (2005) Large-scale analysis of the human ubiquitin-related proteome. Proteomics 5: 4145-4151.

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

69. Mayor T, Lipford JR, Graumann J, Smith GT, Deshaies RJ (2005) Analysis of polyubiquitin conjugates reveals that the Rpn10 substrate receptor contributes to the turnover of multiple proteasome targets. Mol Cell Proteomics 4: 741-751.

70. Mogk A, Schmidt R, Bukau B (2007) The N-end rule pathway for regulated proteolysis: prokaryotic and eukaryotic strategies. Trends Cell Biol 17: 165-172.

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

72. Mukhopadhyay D, Riezman H (2007) Proteasome-independent functions of ubiquitin in endocytosis and signaling. Science 315: 201–205.

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

74. Murantani M, Tansey WP (2003) How the ubiquitin–proteasome system controls transcription. Nat Rev Mol Biol 4: 192-201.

75. Muratani M, Kung C, Shokat KM, Tansey WP (2005) The F box protein Dsg1/Mdm30 is a transcriptional coactivator that stimulates Gal4 turnover and cotranscriptional mRNA processing. Cell 120: 887-899.

76. Murray AW, Kirschner MW (1989) Cyclin synthesis drives the early embryonic cell cycle. Nature 339: 275-280.

77. Newton K, Matsumoto ML, Wertz IE, Kirkpatrick DS, Lill JR et al., (2008) Ubiquitin chain editing revealed by polyubiquitin linkage-specific antibodies. Cell 134: 668-678.

78. Ota K, Kito K, Okada S, Ito T (2008) A proteomic screen reveals the mitochondrial outer membrane protein Mdm34p as an essential target of the F-box protein Mdm30p. Genes Cells 13: 1075-1085.

79. Paiva S, Vieira N, Nondier I, Haguenauer-Tsapis R, Casal M, Grimal-Urban D (2009) Glucose-induced ubiquitylation and endocytosis of the yeast Jen1 transporter: role of Lysine 63-linked ubiquitin chains. J Biol Chem 284:19228-19236.

80. Panse VG, Hardeland U, Werner T, Kuster B, Hurt E (2004) A proteome-wide approach identifies sumoylated substrate proteins in yeast. J Biol Chem 279: 41346-41351.

81. Peng J, Schwartz D, Elias JE, Thoreen CC, Cheng D, Marsischky G, Roelofs J, Finley D, Gygi SP (2003) A proteomics approach to understanding protein ubiquitination. Nat Biotechnol 21: 921-926.

82. Petrucelli L, Dickson D, Kehoe K, Taylor J, Snyder H, Grover A, De Lucia M, McGowan E, Lewis J, Prihar G, Kim J, Dillmann WH, Browne SE, Hall A, Voellmy R,

Tsuboi Y, Dawson TM, Wolozin B, Hardy J, Hutton M (2004) CHIP and Hsp70 regulate tau ubiquitination, degradation and aggregation. Hum Mol Genet 13: 703-714.

83. Pfleger CM, Kirschner MW (2000) The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. Genes Dev 14: 655-665.

84. Pickart CM (2001) Mechanisms underlying ubiquitination. Annu Rev Biochem 70: 503-533.

85. Pridgeon JW, Geetha T, Wooten MW (2003) A method to identify p62's UBA domain interacting proteins. Biol Proced Online 5: 228-237.

86. Rodriguez MS, Dargemont C, Hay RT, (2001) SUMO-1 conjugation *in vivo* requires both a consensus modification motif and nuclear targeting. J Biol Chem 276: 12654-12659.

87. Rogers S, Wells R, Rechsteiner M (1986) Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science 234: 364-368.

88. Saeki Y, Kudo T, Sone T, Kikuchi Y, Yokosawa H, Toh-e A, Tanaka K (2009) Lysine 63-linked polyubiquitin chain may serve as a targeting signal for the 26S proteasome. EMBO J. 28: 359-371.

89. Sampson DA, Wang M, Matunis MJ, (2001) The small ubiquitin-like modifier-1 (SUMO-1) consensus sequence mediates Ubc9 binding and is essential for SUMO-1 modification. J Biol Chem 276: 21664 -21669.

90. Sato K, Hayami R, Wu W, Nishikawa T, Nishikawa H, Okuda Y, Ogata H, Fukuda M, Ohta T (2004) Nucleophosmin/B23 is a candidate substrate for the BRCA1-BARD1 ubiquitin ligase. J Biol Chem 279: 30919-30922.

91. Schlesinger MJ, Bond U (1987) Ubiquitin genes. Oxf Surv Euk Genes 4: 77-91.

92. Seibenhener ML, Babu JR, Geetha T, Wong HC, Krishna NR, Wooten MW (2004) Sequestosome 1/P62 is a polyubiquitin chain binding protein involved in ubiquitin proteasome degradation. Mol Cell Biol 24: 8055-8068.

93. Semenza GL (2000) HIF-1 and human disease: one highly involved factor. Genes Dev 14: 1983-1991.

94. Shimura H, Schwartz D, Gygi SP, Kosik KS (2004) CHIP–Hsc70 complex ubiquitinates phosphorylated tau and enhances cell survival. J Biol Chem 279: 4869-4876.

95. Spence J, Sadis S, Haas AL, Finley D (1995) A ubiquitin mutant with specific defects in DNA repair and multiubiquitination. Mol Cell Biol 15: 1265-1273.

96. Starita LM, Machida Y, Sankaran S, Elias JE, Griffin K, Schlegel BP, Gygi SP, Parvin JD (2004) BRCA1-dependent ubiquitination of gamma-tubulin regulates centrosome number. Mol Cell Biol 24: 8457-8466.

97. Stewart E, Kobayashi H, Harrison D, Hunt T (1994) Destruction of *Xenopus* cyclins A and B2, but not B1, requires binding to p34cdc2. EMBO J 13: 584-594.

98. Sun L, Chen ZJ (2004) The novel functions of ubiquitination in signaling. Curr Opin Cell Biol 16: 119-126.

99. Treir M, Staszewski LM, Bohmann D (1994) Ubiquitin-dependent c-Jun degradation in vivo is mediated by the delta domain. Cell 78: 787-798.

100. Tsirigotis M, Thurig S, Dubé M, Vanderhyden BC, Zhang M, Gray DA (2001) Analysis of ubiquitination in vivo using a transgenic mouse model. Biotechniques 31: 120-6, 128, 130.

101. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403: 623-627.

102. Vasilescu J, Smith JC, Ethier M, Figeys D (2005) Proteomic analysis of ubiquitinated proteins from human MCF-7 breast cancer cells by immunoaffinity purification and mass spectrometry. J Proteome Res 4: 2192-2200.

103. Vertegaal AC, Ogg SC, Jaffray E, Rodriguez MS, Hay RT, Andersen JS, Mann M, Lamond AI (2004) A proteomic study of SUMO-2 target proteins. J Biol Chem 279: 33791-33798.

104. Walters KJ, Lech PJ, Goh AM, Wang Q, Howley PM (2003) DNA-repair protein hHR23a alters its protein structure upon binding proteasomal subunit S5a. Proc Natl Acad Sci USA 100: 12694-12699.

105. Wang H, Matssuzawa A, Brown S, Zhou J, Guy C, Tseng P, Forbes K et al., (2008) Analysis of nondegradative protein ubiquitylation with a monoclonal antibody specific for Lysine-63-linked polyubiquitin. Proc Natl Acad Sci USA 105: 20197-20202.

106. Wilkinson CR, Seeger M, Hartmann-Petersen R, Stone M, Wallace M, Semple C, Gordon C (2001) Proteins containing the UBA domain are able to bind to multi-ubiquitin chains. Nature Cell Biol 3: 939-943.

107. Wilkinson KD (1997) Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. FASEB J 11: 1245-1256.

108. Willems A, Lanker S, Patton EE, Craig AL, Nason TF, Mathias N, Kobayashi R, Wittenberg C, Tyers M (1996) Cdc53 targets phosphorylated G1 cyclins for degradation by the ubiquitin proteolytic pathway. Cell 86: 453-463.

109. Wohlschlegel JA, Johnson ES, Reed SI, Yates III JR (2004) Global analysis of protein sumoylation in *Saccharomyces cerevisiae*. J Biol Chem 279: 45662-45668.

110. Wolfler A, Irandoust M, Meenhuis A, Gits J, Onno R, Touw IP (2009) Site-specific ubiquitination determines lysosomal sorting and signal attenuation of the granulocyte colony-specific stimulating factor receptor. Traffic 10:1168-1179.

111. Won KA, Reed SI (1996) Activation of cyclin E/CDK2 is coupled to site-specific autophosphorylation and ubiquitin-dependent degradation of cyclin E. EMBO J 15: 4182-4193.

112. Wooten MW, Geetha T, Seibenhener ML, Babu JR, Diaz-Meco MT, Moscat J (2005) The p62 scaffold regulates nerve growth factor-induced NF-kappaB activation by influencing TRAF6 polyubiquitination. J Biol Chem 280: 35625-35629.

113. Wooten MW, Hu X, Babu JR, Seibenhener ML, Geetha T, Paine MG, Wooten MC (2006) Signaling, Polyubiquitination, Trafficking, and Inclusions: Sequestosome 1/p62's Role in Neurodegenerative Disease. J Biomed Biotechnol 3: 62079-62092.

114. Wooten MW, Seibenhener ML, Mamidipudi V, Diaz-Meco MT, Barker PA, Moscat J (2001) The atypical protein kinase C-interacting protein p62 is a scaffold for NF-kappaB activation by nerve growth factor. J Biol Chem 276: 7709-7712.

115. Xu J, He Y, Qiang B, Yuan J, Peng X, Pan XM (2008) A novel method for high accuracy sumoylation site prediction from protein sequences. BMC Bioinformatics 9:8.

116. Xu L, Wei Y, Reboul Y, Vaglio P, Shin TH, Vidal M, Elledge SJ, Harper JW (2003) BTB proteins are substrate-specific adaptors in an SCF-like modular ubiquitin ligase containing CUL-3. Nature 425: 316-321.

117. Xu P, Duong D, Seyfriend N, Cheng D, Xie Y, Robert J, Rush J, Hochstrasser M, Finley M, Peng J. (2009) Quantitative proteomics reveals the function of unconvential ubiquitin chains in proteasomal degradation. Cell 137:133-145.

118. Xue Y, Zhou F, Fu C, Xu Y, Yao X (2006) SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res 34: W254-257.

119. Yaglom J, Linskens MH, Sadis S, Rubin DM, Futcher B, Finley D (1995) p34Cdc28-mediated control of Cln3 cyclin degradation. Mol Cell Biol 15: 731-741.

120. Yeh ET, Gong L, Kamitani T (2000) Ubiquitin-like proteins: new wines in new bottles. Gene 248: 1-14.

121. Yen HC, Elledge SJ (2008) Identification of SCF ubiquitin ligase substrates by global protein stability profiling. Science 322: 923-929.

122. Yoshida Y (2003) A novel role for N-glycans in the ERAD system. J Biochem 134: 183-190.

123. Yoshida Y, Chiba T, Tokunaga F, Kawasaki H, Iwai K, Suzuki T, Ito Y, Matsuoka K, Yoshida M, Tanaka K, Tai T (2002) E3 ubiquitin ligase that recognizes sugar chains. Nature 418: 438-442.

124. Zenker M, Mayerle J, Lerch MM, Tagariello A, Zerres K, Durie PR, Beier M, Hülskamp G, Guzman C, Rehder H, Beemer FA, Hamel B, Vanlieferinghen P, Gershoni-Baruch R, Vieira MW, Dumic M, Auslender R, Gil-da-Silva-Lopes VL, Steinlicht S, Rauh M, Shalev SA, Thiel C, Ekici AB, Winterpacht A, Kwon YT, Varshavsky A, Reis A (2005) Deficiency of UBR1, a ubiquitin ligase of the N-end rule pathway, causes pancreatic dysfunction, malformations and mental retardation (Johanson-Blizzard syndrome) Nat Genet 37: 1345-1350.

125. Zhao Y, Kwon SW, Anselmo A, Kaur K, White M (2004) A. Broad spectrum identification of cellular small ubiquitin-related modifier (SUMO) substrate proteins. J Biol Chem 279: 20999-21002.

126. Zhou F, Xue Y, Lu H, Chen G, Yao X (2005) A genome-wide analysis of sumoylation-related biological processes and functions in human nucleus. FEBS Lett 579: 3369-3375.

127. Zhou W, Ryan JJ, Zhou H (2004) Global analyses of sumoylated proteins in *Saccharomyces cerevisiae*. Induction of protein sumoylation by cellular stresses. J Biol Chem 279: 32262-32268.

# CHAPTER 2. IDENTIFICATION OF A CONSENSUS SITE FOR TRAF6/P62 POLYUBIQUITINATION

## ABSTRACT

Tumor necrosis factor receptor-associated factor 6 (TRAF6) is an ubiquitin ligase that regulates a diverse array of physiological processes *via* forming Lys-63 linked polyubiquitin chains. In this study, the Lysine selection process for TRAF6/p62 ubiquitination was examined. The protein sequence of two characterized TRAF6/p62 substrates, NRIF and TrkA, revealed a conserved consensus pattern for the ubiquitination site of these two TRAF6 substrates. The consensus pattern established in the verified substrates was common to the other Trk receptor family members, TrkB and TrkC. Interestingly, Lysine 811 in TrkB was selected for ubiquitination, and mutation of Lysine 811 diminished the formation of TRAF6/p62 complex that is necessary for effective ubiquitination. Moreover, downstream signaling was affected upon binding of BDNF to the mutant TrkB receptor. These findings reveal a possible selection process for targeting a specific Lysine residue by a single E3 ligase and underscore the role of the scaffold, p62, in this process.

*****Published in Biochem Biophys Res Commun. 2008 Jul 4; 371 (3):521-4. Epub 2008 May 5

## INTRODUCTION

Many adaptors have been identified in studies of other neuronal tyrosine kinases that may also prove to function in Trk receptor-mediated signaling (Grimm et al., 2001). Cytoplasmic protein p62 was identified as an interacting partner of atypical protein kinase C (PKC) (Sanchez et al., 1998) and has been shown to contain several protein-protein interacting modules that enable the protein to serve as a scaffold for activation of the transcription factor NF-κB (Moscat et al., 2007). The multidomain protein structure of p62 is suggestive of diverse protein-protein interactions and its link in cellular functions. The functional motifs in p62 include a Phox and Bem1p domain (PB1) domain that embeds an octicosapeptide Phox, Cdc and the atypical PKC-interaction domain (AID) (OPCA) motif, a ZZ zinc finger, a binding site for Tumor necrosis factor Receptor-Associated Factor 6 (TRAF6), two PEST sequences, and an Ubiquitin-associated (UBA) domain (Geetha T. and Wooten MW, 2002). The C-terminal ubiquitin-associated domain (UBA) was discovered to bind non-covalently to ubiquitin (Mueller et al., 2002). In vitro binding studies have unveiled p62 as a unique ubiquitin-binding protein, which binds polyubiquitin non-covalently through its C-terminus (Seibenhener et al., 2004).

Ubiquitination of eukaryotic proteins regulates a broad range of cellular processes. E3 Ub ligases are known to interact with specific substrates either directly or through adaptor proteins. In this regard, p62 has been shown to act as an adaptor and interacts with the TRAF domain of

TRAF6, resulting in its auto-activation (Wooten et al., 2001 and Wooten et al., 2005). Recent findings from have revealed that both TrkA (Geetha et al., 2005) and the neurotrophin receptor interacting factor (NRIF) (Geetha et al., 2005) are K63- polyubiquitinated by the TRAF6/p62 complex. Mutation analyses of these proteins identified a single acceptor Lysine residue that serves as the recognition site for polyubiquitination. TRAF6 possesses a RING finger domain that is responsible for its E3 ligase activity (Rothe et al., 1994). The E3 ligase binds its substrates through its RING domain, which then mediates polyubiquitination of target proteins. TRAF6, together with E2 UBc1/Uve1A, functions as an E3 ligase to mediate the synthesis of K63 linked polyUb chains (Deng et al., 2000).

There are only a few other reports on TRAF6-mediated polyubiquitination that include TRAF6 auto-ubiquitination (Lamothe et al., 2007), NEMO (Lamothe et al., 2007), TAB2 and TAB3 (Ishitani et al., 2003). High substrate specificity of the E3-ubiquitin ligase ensures correct transmission of signals. Yet, little is known about how the substrates are recognized by E3 Ub ligases; nor how site-specific ubiquitination is achieved, and more specifically, why one Lysine may be preferred over the other. In the current study, I investigated this selection process. Close examination of the protein sequence of the verified TRAF6/p62 substrates revealed a consensus pattern therein. This sequence was then used to screen the protein sequence of the other members of the family of Trk receptor proteins. Employing similar bioinformatics predictions a primary ubiquitination site in TrkB and predicted site in TrkC was identified.

## MATERIALS AND METHODS

*Antibodies.* The mouse ubiquitin, HA and p62, rabbit Trk (C-14), HA, and TRAF6 antibodies were purchased from Santa Cruz Biotechnology, La Jolla, CA. Phospho- and nonphospho-MAPK antibodies were purchased from New England Biolabs, and rabbit antibody to phospho Akt (Ser 473), and non-phospho Akt were obtained from Cell Signaling (Beverly, MA). 2.5 S nerve growth factor (NGF), BDNF and NT3 were purchased from Bioproducts for Science (Indianapolis, IN).

*Cell Culture.* Human embryonic kidney (HEK) 293 and nnr5 cells were grown as previously described [8]. HEK293 cells were transfected with the calcium phosphate method by using a Mammalian Cell Transfection Kit (Specialty Media), and nnr5 cells were transfected by using LipofectAMINE 2000 (Invitrogen Life Technologies). The cells were lysed with Triton lysis buffer to detect protein-protein interactions (50 mM Tris-HCl [pH 7.5], 150 mM NaCl, 10 mM NaF, 0.5% Triton X-100, 1 mM $Na_3VO_4$, 1 mM phenylmethylsulfonyl fluoride, and 2 μg/ml leupeptin and aprotinin) or SDS lysis buffer to detect covalent interaction of ubiquitin and TrkA (Triton lysis buffer containing 1% SDS) [6]. Protein was estimated by Bradford procedure (Bio-Rad) and with bovine serum albumin (BSA) as a standard for all samples except those containing SDS, which were estimated by DC assay (Bio-Rad).

*Immunoprecipitation and Western Blotting Analysis.* Cell lysates (1 mg) were diluted in lysis buffer and incubated with 4 μg of primary antibody at 4°C for 3 hr. The immunoprecipitates were collected with agarose-coupled secondary antibody for 2 hr at 4°C and then were washed three times with lysis buffer. The samples were boiled in sodium dodecyl-sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) sample buffer and resolved on gels, transferred onto nitrocellulose membranes, and analyzed by Western blotting with the appropriate antibodies. The samples were separated by 7.5% SDS-PAGE and probed with ubiquitin or TrkA antibodies.

*Site-Directed Mutagenesis.* All primers were obtained from Integrated DNA Technologies, Inc. Coralville, IA and were used without further purification. The forward 32-base primer used to generate the mutant was 5' CTTCAGAACTTGGCG**AGG**GCGTCGCCCGTCTAC 3' and the reverse primer was 3' GTAGACGGGCGACGC**CCT**CGCCAAGTTCTGAAG 5'. QuickChange II XL Site-Directed Mutagenesis Kit was used according to the manufacturer's standard protocol (Stratagene, La Jolla, CA) to mutate A → G at position 3096 resulting in K → R amino acid change in rat TrkB protein sequence (NM_012731.1). The presence of the correct mutation and the absence of PCR-derived alterations to the coding sequence were confirmed by completely sequencing of the mutant receptor construct.

## RESULTS AND DISCUSSION

*Conserved sequences flanking the TRAF6/p62 ubiquitin acceptor site*

Two independent reports have identified TrkA (Geetha et al., 2005) and NRIF (Geetha et al., 2005) as TRAF6/p62 substrates. Moreover, the specific Lysine residue in both these proteins that serve as the ubiquitin acceptor site was identified. Therefore, in an effort to examine similarities in the Lysine selection process for substrate ubiquitination by TRAF6/p62 the protein sequence of TrkA and NRIF was examined. A similarity in the sequences between these two proteins around their primary ubiquitination site revealed a conserved pattern based on chemical properties of the amino acids of the flanking residues at the acceptor Lysine. The consensus pattern observed was [─ (hydrophobic) ─ k ─ (hydrophobic) ─ x ─ x ─ (hydrophobic) ─ (polar) ─ (hydrophobic) ─ (polar) ─ (hydrophobic) -] where k is the ubiquitinated Lysine residue and x any other amino acid (Fig. 1).

Both TrkB and TrkC were examined to determine if this consensus pattern existed in the other members of the Trk family, since both TrkB and TrkC have been reported to be ubiquitinated by TRAF6/p62 (Geetha et al., 2005). Interestingly, Lysines at 811 in TrkB and 602 and 815 in TrkC possessed a similar pattern in their flanking amino acids homologous to the sequence observed in TrkA and NRIF (Fig. 1). Therefore, I hypothesized that these Lysines might act as primary ubiquitin acceptor sites. To test this hypothesis, I focused on the TrkB receptor, since it possessed only one putative ubiquitin acceptor site at K811. In order to test the

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **TrkA_rat** -- gkgsglqghi | G | K | G | S | G | L | Q | G | H | I | K485 |
| **NRIF_mouse**-- vkfedvslf | V | K | F | E | D | V | S | L | T | F | K19 |
| Consensus pattern | * | K | * | X | X | * | ! | * | ! | * | |
| *Putative sites* | | | | | | | | | | | |
| **TrkB_rat** -- akaspvyldi | A | K | A | S | P | V | Y | L | D | I | K811 |
| **TrkC_rat** --vkfygvcgdp | V | K | F | Y | G | V | C | G | D | P | K602 |
| **TrkC_rat --** gkatpiyldi | G | K | A | T | P | I | V | L | D | I | K815 |

**Figure 1.** Conserved sequences flanking the TRAF6/p62 ubiquitin acceptor site. An alignment of TRAF6/p62 ubiquitination acceptor site in NRIF and TrkA shown here with maximum number of matches. Amino acids of the same typed are marked as (*) hydrophobic; (!) polar, (x) any amino acid residue and (k) the acceptor Lysine residue.

possibility that this Lysine was a putative ubiquitin acceptor site, I replaced Lysine at 811 with Arginine (K811R) using site-directed mutagenesis and generated a mutant receptor. The mutant was verified by sequencing. In addition, absence of other mutations was verified by sequencing the entire TrkB cDNA. Secondary structure analysis of protein sequence at an online protein structure prediction server PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) revealed that the ubiquitinated Lysines, K485 of TrkA and K19 of NRIF assumed a coiled-coil motif and K811 in TrkB was likewise predicted to be in a coiled-coil region.

### *TRAF6 ubiquitinates TrkB at Lysine 811*

To check for efficient detection of ubiquitinated wild-type TrkB, HEK293 cells were co-transfected with the HA-tagged TrkB (Wild-type (WT) and Mutant) and the His/myc-tagged Ubiquitin. As control, ubiquitination of WT-TrkA and K485R TrkA mutant (Geetha et al., 2005) was examined. Maximum polyubiquitination of Trk receptors has been observed after 15 min treatment with neurotrophins (Geetha and Wooten 2003). Post-transfection HEK cells were treated with their respective neurotrophin, NGF or BDNF, for 15 min and the extent of receptor ubiquitination was determined by immunoprecipitation with Trk antibody and Western blotting with anti-ubiquitin (Fig. 2, upper panel). TrkA was polyubiquitinated upon addition of NGF and ubiquitination was significantly diminished by mutating K485R (Fig. 2, compare lanes 3 and 5). Likewise, TrkB was polyubiquitinated upon addition of BDNF, while mutation at K811R significantly impaired receptor ubiquitination (Fig. 2, compare lanes 7 and 9). In addition, a fraction of lysate was blotted with Trk antibody to verify the expression levels of all the

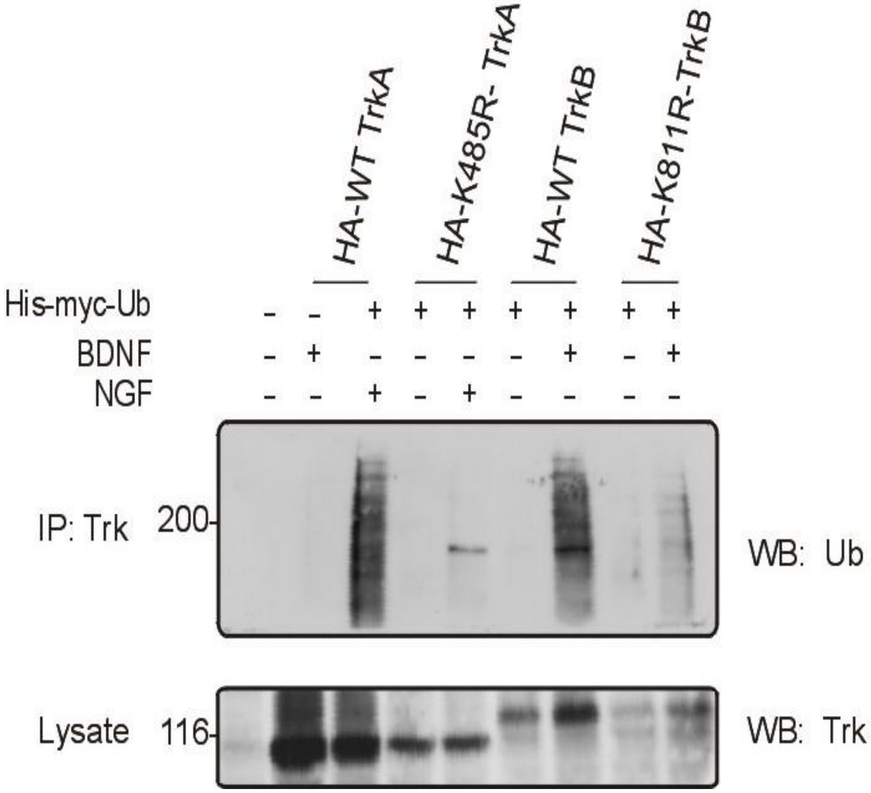

HA-WT TrkA
HA-K485R- TrkA
HA-WT TrkB
HA-K811R-TrkB
His-myc-Ub - - + + + + + + +
BDNF - + - - - - + - +
NGF - - + - + - - - -
IP: Trk
200-
WB: Ub
Lysate
116-
WB: Trk

**Figure 2.** TrkB is ubiquitinated at Lysine 811. HEK cells were transfected with either WT-TrkA, K485R-TrkA, WT-TrkB or K118R-TrkB along with His/Myc-tagged ubiquitin constructs. The cells were treated with or without, NGF for TrkA; BDNF for TrkB for 15 min. The cells were then lysed with SDS lysis buffer and the extent of ubiquitination was determined by immunoprecipitating the lysate with Trk antibody and Western blotting with anti-ubiquitin (upper panel). As a control, a fraction of lysate (50 μg) was blotted with anti-Trk (lower panel). This experiment was replicated three independent times with similar results.

constructs (Fig. 2, lower panel). Consistently diminished expression of TrkB was observed when K811 was mutated to R, suggesting that this Lysine may regulate turnover of the protein. K811 is a primary ubiquitination site, however, a residual amount of polyubiquitin signal was observed on the blot, which might be due to the presence of an additional Lysine residue(s) that is also ubiquitinated. Altogether, these results demonstrate that K811 is a preferential/primary ubiquitin acceptor site in TrkB.

### *Mutation impairs p62's ability to link TrkB to TRAF6*

P62 serves as an adaptor bridge to recruit TRAF6 through its TRAF6 binding site (Moscat et al., 2007; Wooten et al., 2005; Geetha et al., 2005). Therefore, studies were undertaken to examine whether mutation at the primary ubiquitination site in the TrkB receptor impairs formation of a TRAF6/p62 signaling complex. HEK cells were transfected with WT-TrkA, WT-TrkB or their point mutants K485R-TrkA and K811R-TrkB followed by treatment with neurotrophins, either NGF or BDNF, for 15 min to attain maximum polyubiquitination. The cell extracts were immunoprecipitated with Trk antibody and immunoblotted with Trk antibody as control, and TRAF6 and p62 antibody to examine their presence in the complex (Fig. 3). TRAF6 was detected only in lysates recovered from stimulated cells expressing WT receptors (Fig. 3), along with the p62 adaptor (Fig. 3). TRAF6 and p62 were absent in lysates recovered from cells expressing mutant Trk recetpors. These results reveal that mutating the primary ubiquitin acceptor site in either TrkA or TrkB disrupts the interaction between p62, TRAF6 and the Trk receptors (Fig. 3).

### *Receptor ubiquitination implicated in regulation of TrkB downstream signaling*

Ligand binding induces Trk receptors to initiate autophosphorylation. These phosphorylated residues later serve as sites for additional effector factors, and enzymes to bind and propagate the signal downstream. This leads to rapid and sustained activation of various signaling pathways, including the Ras/MAPK pathway and the AKT pathway (Sudo et al., 2000). The ability of NGF and BDNF to stimulate downstream MAPK and AKT signaling as compared to their mutant counterparts was examined. HA-tagged WT-TrkA, WT-TrkB, or their mutants were transfected in nnr5 cells and treated with either NGF or BDNF for 15 min. The lysates recovered from neurotrophin-treated cells were blotted with phospho-MAPK and phospho-AKT antibody, stripped and reprobed with non-phospho antibodies to each protein (Fig. 4). NGF-induced MAPK and AKT activation in the cells expressing WT receptors was impaired in cells expressing mutant TrkA. BDNF had no effect on MAPK activation in the cells expressing the mutant TrkB receptor. However, mutation of K811R in TrkB induced hyper-activation of ATK. This suggests that despite high degree of sequence similarity and broadly overlapping signaling pathways, there still exits divergent signaling response.

## DISCUSSION

Herein I reveal a conserved motif that serves as a recognition determinant for TRAF6/p62 enzyme complex. Ubiquitination is the second most common post-translational modification and is highly conserved in eukaryotes. The choice of Lysine is an important decision as it determines the fate of the protein (Weissman M. 2001). Analysis of the available data on ubiquitination sites

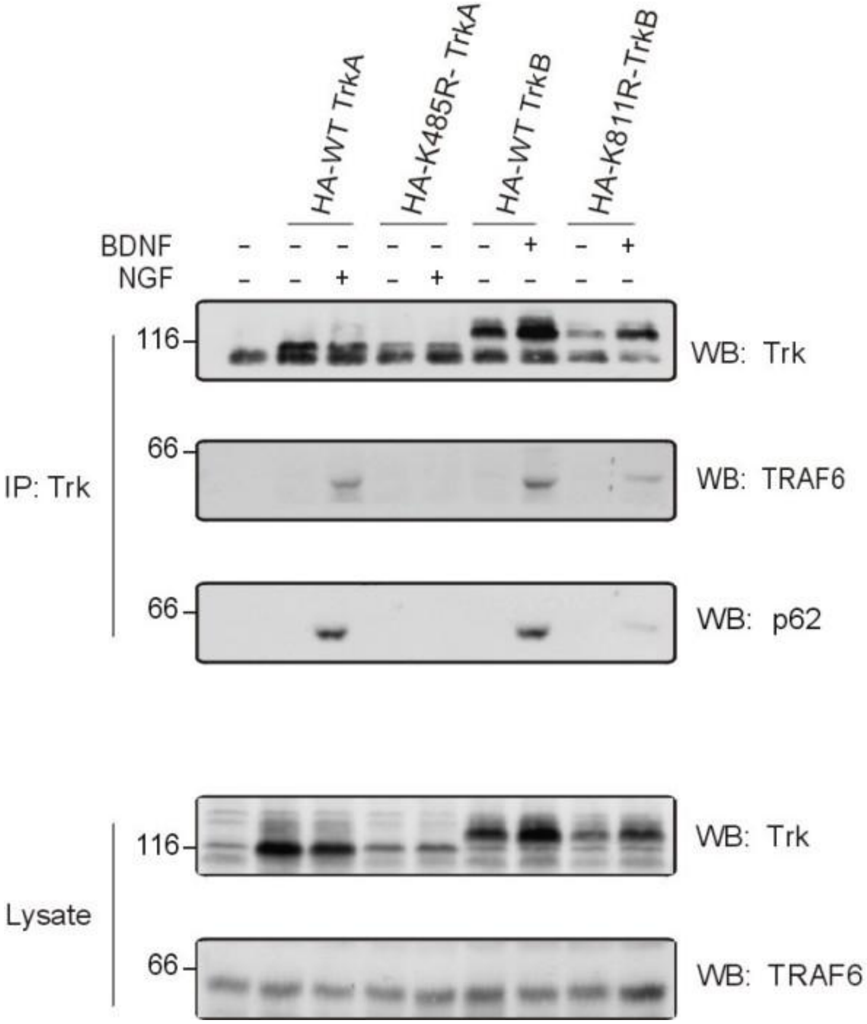

HA-WT TrkA
HA-K485R- TrkA
HA-WT TrkB
HA-K811R-TrkB
BDNF - - - - - - + - +
NGF - - + - + - - - -
116
WB: Trk
66
WB: TRAF6
IP: Trk
66
WB: p62
116
WB: Trk
Lysate
66
WB: TRAF6

**Figure 3.** Mutation of Trk receptors impairs interaction with TRAF6/p62. HEK cells were transfected with WT-TrkA, WT-TrkB or their point mutants K485R-TrkA and K811R-TrkB. The cells were stimulated with or without NGF and BDNF for 15 min. The cells were then lysed in Triton lysis buffer and the cell lysate was immunoprecipitated with Trk antibody and Western blotted with Trk, TRAF6 or p62 antibody. As a control, a fraction of lysate (50 μg) was blotted with anti-Trk and anti-TRAF6 antibody. This experiment was replicated three independent times with similar results.

in yeast showed clear preference for ubiquitination based on structure-function relationship (Catic et al., 2004). Some structural preferences exist for ubiquitin ligation of the targeted proteins such as preferred choice of Lysines in α–helices, and then for easily accessible Lysines in the loop regions. This findings add to the growing list that indicates a bias towards a consensus sequence motif for ubiquitination by a given E3 (Petroski *et al.,* 2003; Ju et al., 2006; Galluzzi et al., 2001; Wu et al., 2003; Scherer et al., 1995; Kumar et al., 2003; Kumar et al., 2004). Moreover, it appears that the E3 targets an accessible surface residue providing the selection process with a conformational recognition mechanism. The TRAF6-p62 signaling complex leads to autoactivation of TRAF6 (Wooten et al., 2005). The scaffold, p62, then recruits the substrate enabling the E3 to scan for the easily accessible Lysine residues in the loops and helical structures on the surface of the substrate resulting in polyubiquitination at a specific Lysine, if the flanking residues fit the consensus motif. This report provides a strategy for studying how TRAF6 defines its Lysine specificity and reveals how scaffolds proteins, on which these complex chemical reactions take place, aid in selecting substrates. Further studies will be needed to develop algorithms and an appropriate search strategy to identify this consensus motif in other TRAF6 and/or p62 interacting proteins.

## ACKNOWLEDGEMENT

This study was funded by the National Institutes of Health (NINDS 33661, MWW).

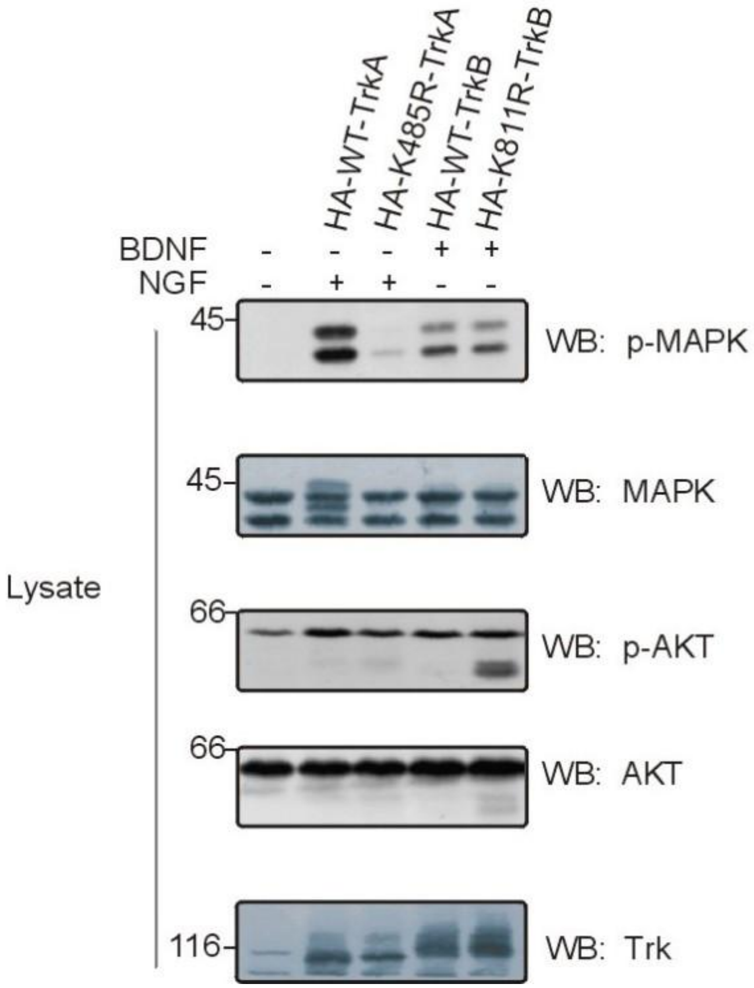

HA-WT-TrkA
HA-K485R-TrkA
HA-WT-TrkB
HA-K811R-TrkB
BDNF - - - + +
NGF - + + - -
45
WB: p-MAPK
45
WB: MAPK
Lysate
66
WB: p-AKT
66
WB: AKT
116
WB: Trk

**Figure 4.** Receptor ubiquitination regulates downstream signaling. HA-tagged WT-TrkA, WT-TrkB, or their mutants were transfected in NNR5 cells and treated with either NGF or BDNF for 15 min. The lysates from transfected cells were blotted with phospho-MAPK and stripped, and reblotted with nonphospho-MAPK antibody as shown. Alternatively, the lysates were also blotted with phospho-AKT and stripped, and reblotted with nonphospho-AKT antibody. The expression of Trk receptors in the lysate was also examined. This experiment was replicated three independent times with similar results.

# REFERENCES

1. Catic C, Collins GM, Church HL (2004) Ploegh, Preferred in vivo ubiquitination sites. Bioinformatics 20: 3302-3307.

2. Deng L, Wang C, Spencer E, Yang L, Braun A, You J, Slaughter C, Pickart C, Chen ZJ (2000) Activation of the IkappaB kinase complex by TRAF6 requires a dimeric ubiquitin-conjugating enzyme complex and a unique polyubiquitin chain. Cell 103: 351-361.

3. Galluzzi L, Paiardini M, Lecomte MC, Magnani M (2001) Identification of the main ubiquitination site in human erythroid alpha-spectrin. FEBS Lett. 489: 254-258.

4. Geetha T, Jiang J, Wooten MW (2005) Lysine 63 polyubiquitination of the nerve growth factor receptor TrkA directs internalization and signaling. Mol Cell. 20: 301-312.

5. Geetha T, Kenchappa RS, Wooten MW, Carter BD (2005) TRAF6-mediated ubiquitination regulates nuclear translocation of NRIF, the p75 receptor interactor. EMBO J. 22: 3859-3868.

6. Geetha T, Wooten MW (2002) Structure and functional properties of the ubiquitin binding protein p62. FEBS Lett. 512: 19-24.

7. Geetha T, Wooten MW (2003) Association of the atypical protein kinase C-interacting protein p62/ZIP with nerve growth factor receptor TrkA regulates receptor trafficking and Erk5 signaling. J Biol Chem. 278: 4730-4739.

8. Grimm J, Sachs M, Britsch S, Di Cesare S, Schwarz-Romond T (2001) Novel p62dok family members, dok-4 and dok-5, are substrates of the c-Ret receptor tyrosine kinase and mediate neuronal differentiation. J Cell Biol. 154: 345– 354.

9. Ishitani T, Takaesu G, Ninomiya-Tsuji J, Shibuya H, Gaynor RB, Matsumoto K (2003) Role of the TAB2-related protein TAB3 in IL-1 and TNF signaling. EMBO J 22: 6277-6288.

10. Ju D, Xie Y (2006) Identification of the preferential ubiquitination site and ubiquitin-dependent degradation signal of Rpn4. J Biol Chem. 281: 10657-10662.

11. Kumar KG, Krolewski JJ, Fuchs SY (2004) Phosphorylation and specific ubiquitin acceptor sites are required for ubiquitination and degradation of the IFNAR1 subunit of type I interferon receptor. J Biol Chem. 279: 46614–46620.

12. Kumar KG, Tang W, Ravindranath AK, Clark WA, Croze E, Fuchs SY (2003) SCF(HOS) ubiquitin ligase mediates the ligand-induced downregulation of the interferon-alpha receptor. EMBO J. 22: 5480–5490.

13. Lamothe B, Besse A, Campos AD, Webster WK, Wu H, Darnay BG (2007) Site-specific Lys-63-linked Tumor Necrosis Factor Receptor-associated Factor 6 Auto-ubiquitination Is a Critical Determinant of IκB Kinase Activation. J Biol Chem. 282: 4102–4112.

14. Moscat JM, Diaz-Meco MT, Wooten MW (2007) Signal integration and diversification through the p62 scaffold protein. Trends Biochem Sci. 32: 95-100.

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

16. Petroski MD, Deshaies RJ (2003) Context of multiubiquitin chain attachment influences the rate of Sic1 degradation Mol Cell. 11: 1435–1444.

17. Rothe M, Wong SC, Henzel WJ, Goeddel DV (1994) A novel family of putative signal transducers associated with the cytoplasmic domain of the 75 kDa tumor necrosis factor receptor. Cell 78: 681–692.

18. Sanchez P, De Carcer G, Sandoval IV J, Moscat, J, Diaz-Meco MT (1998) Localization of Atypical Protein Kinase C Isoforms into Lysosome-Targeted Endosomes through Interaction with p62 Mol Cell Biol. 18: 3069-3080.

19. Scherer DC, Brockman JA, Chen Z, Maniatis T, Ballard DW (1995) Signal-induced degradation of I kappa B alpha requires site-specific ubiquitination. Proc Natl Acad Sci. USA 92: 11259–11263.

20. Seibenhener ML, Babu JR, Geetha T, Wong HC, Krishna NR, Wooten MW (2004) Sequestosome 1/P62 is a polyubiquitin chain binding protein involved in ubiquitin proteasome degradation. Mol Cell Biol. 24: 8055–8068.

21. Sudo T, Maruyama MH, Osada H (2000) P62 functions as a p38 MAP kinase regulator. Biochem Biophys Res Commun. 269: 521-525.

22. Weissman M (2001) Themes and variations on ubiquitylation. Nat Rev Mol Cell Biol. 169-178.

23. Wooten MW, Geetha T, Seibenhener ML, Babu JR, Diaz-Meco MT, Moscat J (2005) The p62 scaffold regulates nerve growth factor-induced NF-kappaB activation by influencing TRAF6 polyubiquitination. J Biol Chem. 280: 35625-35629.

24. Wooten MW, Seibenhener ML, Mamidipudi V, Diaz-Meco MT, Barker PA, Moscat J (2001) The atypical protein kinase C-interacting protein p62 is a scaffold for NF-kappaB activation by nerve growth factor. J Biol Chem. 276: 7709-7712.

25. Wu G, Xu G, Schulman BA, Jeffrey PD, Harper JW, Pavletich NP (2003) Structure of a β-TrCP1-Skp1-β-Catenin Complex Destruction Motif Binding and Lysine Specificity of the SCFβ-TrCP1 Ubiquitin Ligase. Mol Cell. 11: 1445–1456.

# CHAPTER 3. COMPUTATIONAL SEARCH FOR PREFERRED TRAF6/P62 UBIQUITINATION SITES: A TEST OF THE "CODE-HYPOTHESIS"

## ABSTRACT

There are approximately one thousand reported E3 ligases in eukaryotes. The preferred substrates for most of these enzymes remain unknown. Moreover, it remains unclear how among the many Lysines (K) found in an ubiquitinated protein only a few are targeted as *bona fide* ubiquitination sites. Furthermore, cellular E3 ligases and scaffold proteins interact with numerous binding proteins through their multi-domain structures. These interactors could be potential ligase substrates. A new approach is described here to predict ubiquitinated substrates of the TRAF6/p62 complex. I observed that although there was low linear conservation of a single consensus motif at predicted ubiquitinated sites, there is substantial structural and evolutionary conservation of a generalized motif surrounding these predicted sites. Analysis revealed that the identified target sites have structural preferences as well as a dependence on accessibility within the protein molecule.

## INTRODUCTION

E3 protein ligase is the component of the ubiquitin conjugation system that is most directly involved in substrate recognition. There are approximately 617 genes encoding putative Ub E3s which is more than the 518 genes reported for protein kinases (Li et al., 2008). Preferred substrates for most of these enzymes remain unknown. The biological importance of E3s requires understanding the site selection process involved in substrate recognition during ubiquitination. Eukaryotic cells express a single ubiquitin-activating enzyme (E1) that activates free ubiquitin for subsequent transfer to one of approximately 50 ubiquitin-conjugating enzymes (E2) (Willis et al., 2008). Ubiquitin E3 ligases recruit both substrate and activated ubiquitin to mediate the transfer of the ubiquitin molecule to the targeted protein either directly or with the help of E2 enzymes (Liu, 2004). The substrate specificity of the ubiquitination process occurs at the level of the E3 ubiquitin ligases. Large numbers of cellular proteins are known to be ubiquitinated and correspondingly, there are large numbers of E3 ligases with a diverse range of structures.

Ubiquitination is a complex process. Only a few Lysines (K) out of many in a target protein are ubiquitinated. Moreover, ubiquitination is very dynamic. Less than 1% of the cellular proteins are ubiquitinated *in vivo* at any given time. In this regard, our understanding of the ubiquitination process is still in its infancy. A number of *in vivo* and *in vitro* methods have been employed to identify ubiquitinated substrates and their sites, including proteome-scale analyses of the substrates (Peng et al., 2003; Matsumoto et al., 2005; Jeon et al., 2007). All these methods are time-consuming, labor-intensive, and expensive. In addition, they are focused on characterizing the 'ubiquitinated proteome' rather than studying single enzyme substrates. In contrast, computational approaches represent promising alternative methods for identification of ubiquitination sites.

Until recently, no consensus amino acid motif had been reported for a single ligase enzyme (Jadhav et al., 2008). The reported biological specificity seems to be associated with substrate selection. This observation prompted  to hypothesize that there exists an ubiquitination 'language' that encodes specific amino acid patterns in the substrate that is read by E3 ligases. Here, I refer to this language as a "code hypothesis".  Using the code hypothesis as the base, I developed a method to predict putative TRAF6/p62 ubiquitination sites using consensus motif pattern information. To facilitate identification of the consensus motifs within putative substrate proteins, a brute-force motif search algorithm was designed and implemented.

**Code hypothesis**

Independent studies have identified two TRAF6/p62 substrates, tyrosine receptor kinase A (TrkA) (Geetha et al., 2005) and Neurotrophin receptor interacting factor (NRIF) (Geetha et al., 2005) (Fig. 1A). Both of them were K63- polyubiquitinated at their target Lysines. Mutagenesis studies of these proteins verified the acceptor Lysine residue that served as the target site for polyubiquitination. The RING finger domain of TRAF6 ligase is known to be responsible for its catalytic E3 ligase activity (Lamothe, B. et al., 2007). E3 ligase binds its substrates through its RING domain (Deshaies, R. and Joazeiro, C., 2009), which then mediates polyubiquitination of target proteins. UBc1/Uve1A functions as an E2 enzyme that mediates the transfer of activated ubiquitin molecules in this reaction (Geetha. et al., 2005; Geetha et al., 2005). Modular protein p62 provides the platform for the transfer reaction to occur. There are only a few other reports on TRAF6-mediated polyubiquitination, including TRAF6 auto-ubiquitination, NEMO (Lamothe et al., 2007), TAB2 and TAB3 (Ishitani et al., 2003). These reactions, however, have not been shown to require p62 to mediate the modification. Moreover, like TRAF6, there are many reported E3 Ub ligases in the literature, whose potential pool of biological targets are unknown.

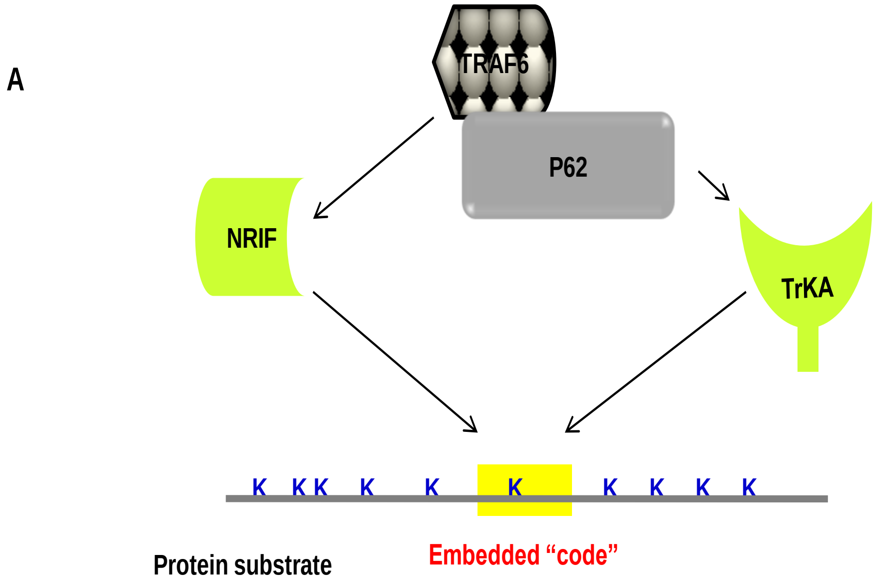
The process of cell signal transduction is dependent on specific protein-protein interactions. Within protein-protein interaction networks, most proteins interact with a few partners. However, a small number of proteins – called 'hubs' – interact with many different partners forming multimeric signaling complexes. These hubs mediate interactions by their modular protein domains that confer specific binding activity to their interacting partners. Protein p62 contains several structural motifs that allow it to act as a hub for protein-protein interactions. These motifs include an acidic interaction domain (AID/ORCA/PC/PB1) that binds the aPKC, a ZZ finger, a binding site for the RING finger protein TRAF6, two PEST sequences, and the UBA domain (Geetha and Wooten 2002). In this work, I focused on the mechanism by which TRAF6, along with p62, recognizes target Lysines on its substrates as ubiquitin acceptors. In the enzyme-substrate model, p62 is suggested to serve as a crucial bridge between enzyme (E3 ligase, TRAF6) and its substrate(s), and provide specificity for enzyme-substrate reactions. Thus, substrate recognition, site selection, and ultimately the ubiquitination reaction, result from the concerted action of the active TRAF6/p62 enzyme complex.

As the starting point for this research, I examined protein sequences of two known TRAF6/p62 substrates. This initial analysis concentrated on target ubiquitination sites selected to optimize my search for any potential consensus motif. Examination of flanking residues surrounding the target Lysine did reveal the presence of a likely consensus motif, which was then used to screen the protein sequences derived from Trk receptor family. Ubiquitination sites in TrkB and TrkC proteins were first identified *in silico* (Jadhav, T., 2008) and then confirmed through site-directed mutagenesis and functional testing. The identified consensus motif was further characterized. The final analysis identified a 10-amino acid long sequence of [-hydrophobic – k – hydrophobic – x – x – hydrophobic - polar1 – hydrophobic - polar2 – hydrophobic -]. The hydrophobic amino acids included Alanine, Leucine, Valine, Methionine, Glycine, Phenylalanine, or Isoleucine. The polar1 amino acids included Glutamine, Tyrosine, Cysteine, or Serine and polar2 included Histidine, Aspartic Acid, or Threonine (Fig. 1B).

## METHODS

### Database preparation

First, to test the hypothesis that TRAF6- and p62- interacting proteins are putative E3 Ub ligase substrates, I developed a protein database. These proteins within this database were divided into two groups, an experimental dataset and a negative dataset. The experimental dataset proteins were further divided into five groups depending on their probability of being a TRAF6/p62 substrate (see Table 1). These ranged from known ubiquitinated substrates with mapped sites to either TRAF6 or p62 interacting proteins. All known TRAF6/p62 substrates with verified ubiquitination (Ub) sites were placed in group I. Group II contained known and tested substrates of TRAF6 E3 ligase whose target Lysine Ub site(s) were not mapped nor identified and their interaction status with p62 unknown.  TRAF6- and p62- interactors identified from various protein-protein interaction databases [HPRD (Prasad et al., 2009), and BioGRID (Breitkreutz et al., 2008) and EntrezGene (Maglott et al., 2005)] formed Groups III and IV,respectively. Finally, Group V comprised of proteins from the insoluble Formic acid (FA) fraction of the brain from p62 knockout mice. The negative dataset contained 54 proteins selected from the NCBI database with no reports of interaction with either TRAF6 or p62 proteins (see Table 1). This dataset was used both for control comparisons and as a test group for the developed algorithm. Proteins in the database were curated for their localization, domain structure and function. In total, 211 protein sequences were examined for the presence of TRAF6/p62 ubiquitination site(s), of which 157 proteins sequences belonged to the experimental dataset and 54 protein sequences to the negative dataset (Table 1).



B

10 AA long consensus motif

– (hydrophobic) – K – (hydrophobic) – X – X – (hydrophobic) – (polar1) – (hydrophobic) – (polar2) – (hydrophobic) –

*where;* K = ubiquitinated lysine, X = any amino acid; hydrophobic = A, V, F, P, M, L, I, G; polar1 = Q,Y,C,S ; polar2 = H,D,T

**Figure 1. A.** Schematic representation of “code hypothesis”. **B.** The refined consensus motif identified in TRAF6/p62 substrates.

### Motif search protocol

Amino acid sequences of the 211 proteins in the database were searched using a brute-force approach. First, I generated a file containing all unique combinations of seven variable positions in the 10 amino acid long target motif (hydrophobic – k – hydrophobic – x – x – hydrophobic – polar1 – hydrophobic – polar2 – hydrophobic). Hydrophobic amino acids included in the motif were Alanine, Phenylalanine, Glycine, Isoleucine, Leucine, Methionine and Valine. The polar1 category contained either Cysteine, Glutamine, Serine or Tyrosine; polar2 amino acids included either Aspartic acid, Histidine or Threonine. Excluding the two positions (x) that could contain any amino acid, a total of 201,684 unique seven position motifs were possible. I employed two computer-based search algorithms to facilitate the screening process for the presence of consensus motifs. The first program, MotifMaker, is a simple looping program. It generated and stored all 201,684 potential motifs. The second program, MotifFinder, implemented a brute-force search algorithm for all possible motif constructs. The analysis started by identifying and counting each ***K*** within the target peptide. Any ***K*** within 8 residues from the carboxyl end was excluded because it would be impossible for it to be a full motif. The motif search then proceeded by temporarily storing the ***K-1, K+1, K+4…K+8*** amino acids for each ***K*** as a character string and comparing this string to each of the 201,684 potential motif patterns. A step-up procedure was used to determine the best fit among the potential motifs. For each ***K***, a counter would be initially set at “zero” matches. The counter would be progressively updated as positive matches between the target string and potential motifs were encountered. The matching motif would then be stored in the computer memory. By searching all possible motif combinations, this approach ensured that the maximum ‘best match’ motif was identified. In motifs that matched at all 7 variable positions, a perfect match was identified. In motifs with less than perfect matches (6, 5, 4,…1), the algorithm ensured that no motif with a greater number of matching locations could be found. The procedure was repeated at each ***K*** within the target peptide until all positions had been searched. Information on the location of each ***K***, the pattern in motifs that matched at all 7 variable positions, a perfect match was identified.

In motifs with less than perfect matches (6, 5, 4,…1), the algorithm ensured that no motif with a greater number of matching locations could be found. The procedure was repeated at each ***K*** within the target peptide until all positions had been searched. Information on the location of each ***K***, the pattern of hits, the amino acid sequence of both the target, best match and the total count of positive hits were collected as an output. Both programs were developed and executed using MATLAB® V6.5 (MathWorks Inc., Natick MA).

**Sequence logos**

Sequence logos for displaying the flanking residue distribution of all Lysines in the datasets were created using the web-based program WebLogo (Schneider et al., 1990; Crooks et al., 2004). The height of each letter in the stack is proportional to its frequency at that position in the consensus motif. Letters were sorted with the most frequent amino acid on top.

**Table 1.** Database proteins classification system and distribution of proteins in each dataset.

| | | | |
|---|---|---|---|
| Experimental dataset | Group I | Verified TRAF6/p62 substrates | 4 |
| | Group II | Predicted TRAF6 substrates | 7 |
| | Group III | TRAF6 interactors | 59 |
| | Group IV | P62 interactors | 37 |
| | Group V | Insoluble Formic acid (FA) fraction proteins in p62 knockout mice | 50 |
| Negative dataset | | Control group with no documented TRAF6 or p62 interaction | 54 |
| Total proteins | | | 11 |

## Statistical analysis

A total of 211 proteins were examined for the consensus TRAF6/p62 motif at the flanking residues of Lysines and were scored for their frequency (Appendix, Table T1). The distribution of frequency hits for the consensus motif between experimental and negative datasets was statistically compared using Chi-square analysis. Kurtosis (Pearson and Hartley, 1972a) and skewness (Pearson and Hartley, 1972b) generated from each empirical distribution were also statistically compared. All calculations are based on the $\chi$2-test with Yates' correction (one degree of freedom).

## Secondary structure prediction

PSIPRED (Jones, 1999; Bryson et al. 2005) was used to predict secondary structures. PSIPRED uses neural networking and searches for homologous proteins with known structures to determine the most likely structure at each residue position. Predictions of disorder regions at the predicted ubiquitinated sites were made using the Metaserver of Disorder (MeDor) (Lieutaud et al., 2008). MeDor collects disorder and secondary structure predictions from servers available on the web and generates a graphical output. The web-based database SMART (Schultz. et al., 1998) was used to predict signaling domains within the protein sequences identified as containing strong motif patterns. The SABLE server was used to predict from sequence secondary structures and solvent accessibilities, with the goal of identifying potential characteristics of predicted Ub sites in terms of structural profiles (Adamczak et al., 2004).

## RESULTS AND DISCUSSION

### Analysis of ubiquitination motif

Results from the motif search analyses revealed a wide range of distribution in amino acids surrounding Lysines in both datasets and with positive hits ranging from 1-7 (where a hit of 7 was perfect hit) in the experimental proteins and 1-6 in the negative dataset. As expected, tests for distributional pattern indicated a strong departure from normal distribution for both datasets. Based on these results, measures of skewness and kurtosis were examined to better understand the pattern of positive hits. Both datasets exhibited positive kurtosis values as reflective of their peaked distribution, leptokurtic. The value of kurtosis for the experimental dataset (2.67) containing substrates/interactors, for frequency bands with positive hits at positions >3 was significantly higher ($p > 0.05$) than that of the negative dataset (2.82). Positive skewness values indicated that the motif hit distributions for both datasets were strongly asymmetric (1.66 for experimental dataset and 1.71 for negative dataset; $p > 0.05$). Furthermore, the most obvious pattern was a substantial shift in the distribution of positive hits in the experimental dataset relative to the results from the negative proteins (Appendix, Figure S1 and S2). Collectively, these results provided critical information regarding the comparison of the experimental and negative protein datasets. First, the similarity between the negative and experimental datasets suggested that the selection of potential interacting proteins for the experimental group did not overtly bias the results. Conversely, the presence of perfect motif matches and more (< 4) positive hits for the experimental group suggested that the perfect motif is associated with known function.

**Statistical profile of the motif hits**

Goodness-of-fit tests were used to examine how well the observed data and expected values derived from the negative and experimental datasets fit, respectively. I investigated whether the distribution of positive hits in the negative dataset conformed to the distribution of positive hits in the experimental dataset. The observed Chi-squared statistic (0.855) exceeded the critical value for the 0.05 probability level. This finding indicated that the observed values from the negative hit distribution differed significantly from that of expected values in the experimental dataset. Specifically, consistent with both visual observation of the distribution patterns and the skewness/kurtosis estimates, a higher proportion of strong positive hits were encountered in the experimental proteins relative to those in the negative dataset (Appendix, Table T7-8).

Next, I sought to find amino acids that play a critical role in ubiquitination site selection, and investigated whether there were preferences for certain amino acids near the target ubiquitinated Lysines. This analysis focused on the well-defined proteins from Group I of the experimental dataset. Notably, when I examined the surrounding residues of the validated ubiquitinated Lysine with amino acids conserved at 7 variable positions in the hypothesized motif (perfect hit), I discovered an enrichment of small residues (G/A) on the either side of the target side and high frequency of Valine at position 4, and Leucine at position 6, and Aspartic acid at position 7 (see Fig. 2A). A closer look at all proteins from the experimental dataset (Groups I through V) with amino acids conserved at 6 positions revealed a similar distribution of amino acids, (see Fig. 2B). When the distribution of amino acids positive hits at 6 positions were compared for both the datasets, the amino acid distribution in the negative dataset of non-interactors proteins was much more indiscriminate around the Lysine residue (see Fig. 2C). However, in all datasets, the target Lysine residue was predominantly surrounded by hydrophobic residues (Glycine/Alanine/Valine/Leucine/Isoleucine).

**Secondary structure prediction**

Because post translational modifications tend to be concentrated within specific structural regions of a protein, I further investigated structural constraints of the predicted Lysines. Only predicted Lysines from highly positive (conserved at 6 or 7 variable sites) motif sites were included in this analysis (Appendix, Table T9 and T10). These Lysines were classified as a high probability group. There were total of 30 proteins in this category, 25 from the experimental dataset and 5 from the negative dataset containing a total of 37 high probability sites. Eight of those 30 proteins had more than one predicted TRAF6/p62 ubiquitination site (Appendix, Table T11). Proteins NRIF, TRKA, TRKB, TRKC, NTRK2, NTRK3 and MBP had perfect match to the hypothesize motif for TRAF6/p62 ubiquitination. GO ontology analysis of these high probability proteins with perfect match reveled that they were involved mainly in membrane bound signaling events (Appendix, Table T12). I sought to incorporate sequence information as well as information from sequence derived structural features of these proteins into the validation process. To do so, four potential structural features of the predicted high probability sites were evaluated: secondary structure, relative distribution within the protein, solvent accessibility, and the intrinsic disorder within the protein domain.

These results indicated that approximately one-half of the predicted ubiquitination sites are predicted to be in loops and disordered regions (Fig.3A). Beta-sheets had the least representation of predicted ubiquitination sites (with 15% sites in experimental and none from negative datasets). The predicted ubiquitinated site was found at a significantly greater rate in the loop regions than in the beta sheets of the protein structure ($P$ = 0.0001). The second most common secondary structure was an alpha-helix (Fig.3A). Alpha helices and loops are usually found on the surface of proteins and are tend to easily accessible for posttranslational modifications. The predicted sites show significantly high occurrence of sites in helices and loops as compared with occurring in beta sheets ($P$ = 0.0001). This was in agreement with previously reported findings on preferred *in vivo* ubiquitination sites in yeast (Catic et al., 2004). The critical position of Lysine 507 of Smad4 was recognized from detailed crystallographic studies of the fully solvent-accessible L3 loop with its side chain protruding from the L3 loop surface to the neighboring space (Morén et al., 2003).

### C-terminal Lysines

The highest possible resolution for investigating structure–function relationships is that of individual residues and their corresponding microenvironments (Wu, S. 2010). To provide information on this aspect of hypothetical high-probability sites, the distribution of predicted Lysines residues with regards to their relative position within the protein sequence was searched. Nearly half (48%) of the motif target Lysines were located near the C-termini of the proteins in the experimental dataset as compared to only 28% in the negative dataset. The remaining predicted sites were evenly distributed (25.8%) at the C-terminus or middle region of the proteins in the experimental dataset. On the contrary, within the negative dataset, most (42%) target Lysines were found in the middle region of the protein (Fig. 3D). This could be either because of false positive prediction of the sites or due to true positive (valid) sites that are buried inside the protein and become exposed when these proteins undergo conformational changes induced by either other posttranslational modifications or protein-protein interactions. This finding is consistent with studies of the TRAF6 substrate, IRF7 that is ubiquitinated at multiple sites both *in vitro* and *in vivo* with the three C-terminal Lysines (positions 444, 446, and 452) essential for activation of IRF7 (Chew et al., 2006; Ning et al., 2008). Similar studies on SUMOylation sites of LEDGF/p75 have shown that K75, K250, and K254 mapped on the N-terminal region located in evolutionarily conserved charge-rich regions, while C-terminal K364 was identified as solvent exposed (Bueno et al., 2010). There were 86 lysines in the N-terminal regions of the proteins in the experimental dataset that were not recognized by the program as they lacked the required 8 amino acids towards the N-terminal end to fit the 10 amino acid long motif condition. Out of these 86 Lysines, there were five instances of di-Lysines and four tri-Lysines with one occurrence of poly-lysine chain of 9 lysines. Negative dataset, on the other hand had 29 N-terminal Lysines, with only one occurrence of di-Lysine in the NCL protein. No specific amino acid distribution pattern was observed surrounding the N-terminal lysines. The downstream Lysines in the di-Lysine sequences have been reported to be preferentially ubiquitinated in the examined yeast ubiquitination sites (Catic et al., 2004).

### Surface accessibility

Recent studies of all post translationally modified proteins documented in Swiss-Prot has shown that most reversible modifications are found on the protein surfaces (Pang et al., 2007). Ubiquitinated Lysines are surface exposed but this information is hidden in the primary sequence of the protein which can be detected by the surface accessibility predictor. To examine this possibility for the data, solvent accessibility of the high probability target Lysines for modification was examined. Solvent accessibility of an individual residue is often classified as “buried" or "exposed" using geometric analysis (geometric similarity in the arrangement of the water molecules around proteins) (Britton et al., 2006) or predictive methods. Prediction of solvent accessibilities revealed 84% of the highly positive motif sites in the experimental dataset and 100% of the negative dataset were exposed on the surface of the protein ($P$ = 0.009), which in a cellular environment, would be easily accessible to the active TRAF6/P62 complex (Fig. 3C). It has been reported that surface accessibility of post-translational modifications is important for protein−protein interactivity (Pang et al., 2007). Moreover, since proteins involved in cellular signaling are predicted to have long disordered regions, surface accessibility prediction was performed on the 30 high probability substrates in the database. The structural environment of TRAF6/p62 predicted sites was assessed to check whether the predicted Lysines sites occurred in ordered or in disordered regions. Structural analysis was conducted using secondary structure, protein domain, and disorder prediction algorithms (Lieutaud et al., 2008). Predicted ubiquitination sites were found to be predominantly located in coils or disordered regions (Fig. 3B and C).

**Compartment specific ubiquitination motif**

Next, to study the subcellular distribution of the predicted TRAF6/p62 ubiquitination substrates compartmentalization of the proteins in both the datasets was examined (Appendix, Table T2-6). Proteins were assigned to cellular compartments based on the literature evidence, curated information in protein databases and GO ontology for protein subcellular localization (Harris et al., 2004).

Localization data of the high probability substrates revealed that relatively few cytosolic proteins predicted to be TRAF6/p62 substrates. However, when the nuclear proteins in the experimental dataset were compared, slightly more substrates (29%) were predicted as compared to 25% composition of nuclear protein in the dataset (Fig. 4). A substantial increase in prediction of substrates was observed for proteins that were integral to membranes in both the experimental and negative datasets ($P = 0.03$). This finding shows that since the consensus motif was based on plasma membrane bound TrkA and nuclear protein NRIF, the two TRAF6/p62 substrates (Geetha et al., 2005 and Geetha et al., 2005), it was biased to predicting membrane bound and nuclear proteins. The consensus motif can be further refined as more substrates ar verified experimentally from various subcellular localizations. Moreover, this study points out the need

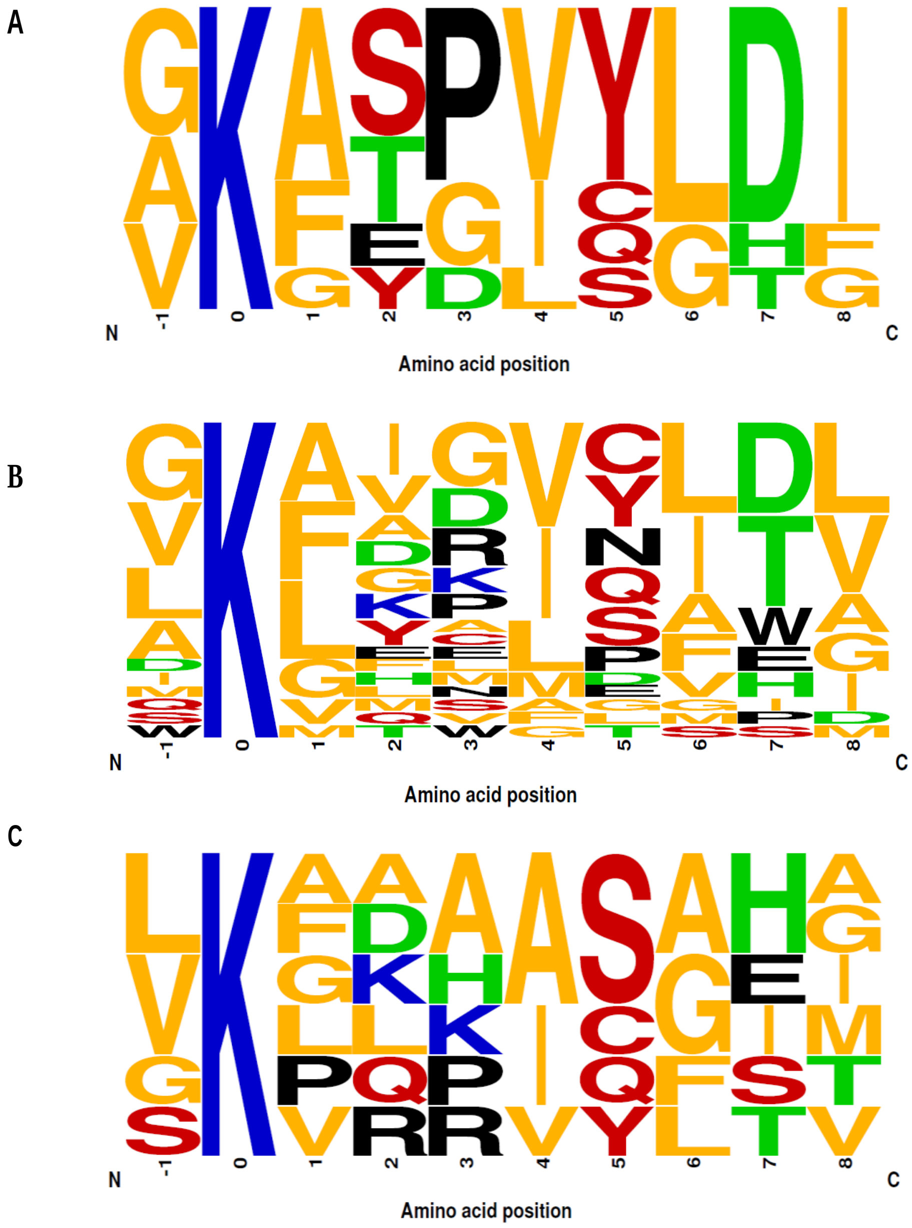

A
B
C
N
C
-1
0
1
2
3
4
5
6
7
8
Amino acid position

**Figure 2.** Frequency distribution of amino acids surrounding Lysines (K) with positive hits.

**A.** at seven variable positions in a ten amino acid long consensus motif in the experimental dataset, **B.** at six variable positions in a ten amino acid long consensus motif in the experimental dataset and **C.** at six variable positions in a ten amino acid long consensus motif in the negative dataset. K (red); AFGILMV (blue); CQSY (green); DHT (orange).

for a prediction system based on individual E3 enzyme systems where the linear recognition motif signature is further enhanced by structural features derived from the overall sequence.

### Sequence conservation

I sought to further validate the biological relevance of my hypothetical ubiquitination motif by examining it in an evolutionary context. There are no examples of proteins where more than one homolog has been investigated for its ubiquitination sites. So high-confidence set of TRAF6/P62 substrates, had the sites with exact match to the consensus motif, were selected for alignment. The proteins with exact match to the consensus motif are TrkA, TrkB, TrkC, NRIF, NTRK2, NTRK3 and MBP. To check for potential evidence of extra evolutionary pressure to conserve the site-specific ubiquitinated lysines, conservation of the predicted sites in these eight proteins was examined across multiple species. These results indicated that a predicted ubiquitination sites were conserved from among six mammalian species (Appendix, Figure S3). This unusually high conservation suggests that the ubiquitination of these sites may be also be conserved in all life forms, although this has still to be proven. A high degree of conservation among proteins that are ubiquitinated also suggests that they may have arisen early in the course of evolution. However, a significant number of ubiquitination sites differ in the ubiquitome and the extent of homology is not uniform because of the high diversity among the proteins. Nevertheless, evidence of conservation does suggest that ubiquitination is in each case indispensable for protein function, which is in turn essential for regulating cellular function. These highly conserved essential ubiquitination events may reflect how the earliest forms of life used protein ubiquitination in specific housekeeping cellular functions. Interestingly, results in this study indicated that although the surrounding sequence regions may diverge, the critical residues remain conserved. Similar whole genome-scale studies have shown that 2683 potential SUMO substrates are conserved between human and mouse based on the pattern recognition and phylogenetic conservation (Zhou, 2005). In another study linear pattern recognition in combination with phylogenetic conservation was first used to discover transcription factor binding sites (Loots, 2007). This finding is similar to results from recent studies on phosphorylation sites that have shown that these sites that demonstrated similar conservation within protein families (Maathuis, 2008) thus pointing at generic regulatory mechanisms which may be conserved across species. This is indicative of the fact that the short length and the rare conservation over long evolutionary distances make linear motifs difficult to find computationally (Neduva, 2005).

## CONLCUSIONS

Conservation of target-specific amino acid sites within a protein is often taken to imply biological importance. To test the generality of this finding, I analyzed the structures of 30 proteins that were predicted to be TRAF6/P62 substrates. A total of 37 predicted TRAF6/p62 ubiquitination sites were identified. It was observed that the predicted ubiquitination sites were biased towards the C-terminal domain of the protein, as previously reported (Chew et al., 2006; Ning et al., 2008). Secondary structure analysis of the predicted sites revealed overall preference for loops and helices. Tertiary structure analyses of investigated proteins revealed that most of the predicted sites are likely to be exposed on the surface of the protein rather than being buried. Although linear conservation of individual amino acids within the consensus motif at the predicted ubiquitinated sites is low, there is a high structural and evolutionary conservation of predicted sites across mammalian species. The high accessibility of ubiquitination sites suggests that they are localized in loops and helices, since these structural elements are usually found at the protein surface. It is well known that the loop regions frequently participate in forming binding sites and active sites of enzymes making them excellent substrates for regulation (Gnad et al., 2007). Beta sheets can be internal to a protein (largely hydrophobic) or on the surface in which case they are amphipathic, with every other amino acid side chain alternating between hydrophobic and hydrophilic nature. Because posttranslational modification sites are predominantly located in rapidly evolving loop regions (Gnad et al., 2007), relaxed evolutionary constraints on loops allow them to evolve rapidly and rather independently from the protein core. Formally, disordered regions are defined as regions within proteins that lack a precise 3D structure and consist of an ensemble of fluctuating, interconverting conformers. These regions have been known to be associated frequently with posttranslational modifications (Fuxreiter et al., 2008). Disorder prediction of linear motifs and their flanking regions for the experimentally characterized examples from the Eukaryotic Linear Motif (ELM) database revealed that short recognitions motifs are embedded in locally unstructured regions (Fuxreiter et al., 2007). Thus, structurally and evolutionarily, the high-confidence set of TRAF6/62 substrates and highly positive motif sites represent a reasonable site for modification by ubiquitin.

In conclusion, a holistic approach to use a combination of sequence motif data and structural determinants along with evolutionary conservation can greatly aid in identification of the substrates and prediction of putative ubiquitination sites. Presence of high amount of plasma membrane proteins in the high probability dataset indicate that the “code hypothesis” can be applied to other E3 ligases for prediction of their substrates taking into account their binding partners, or adaptor molecules.

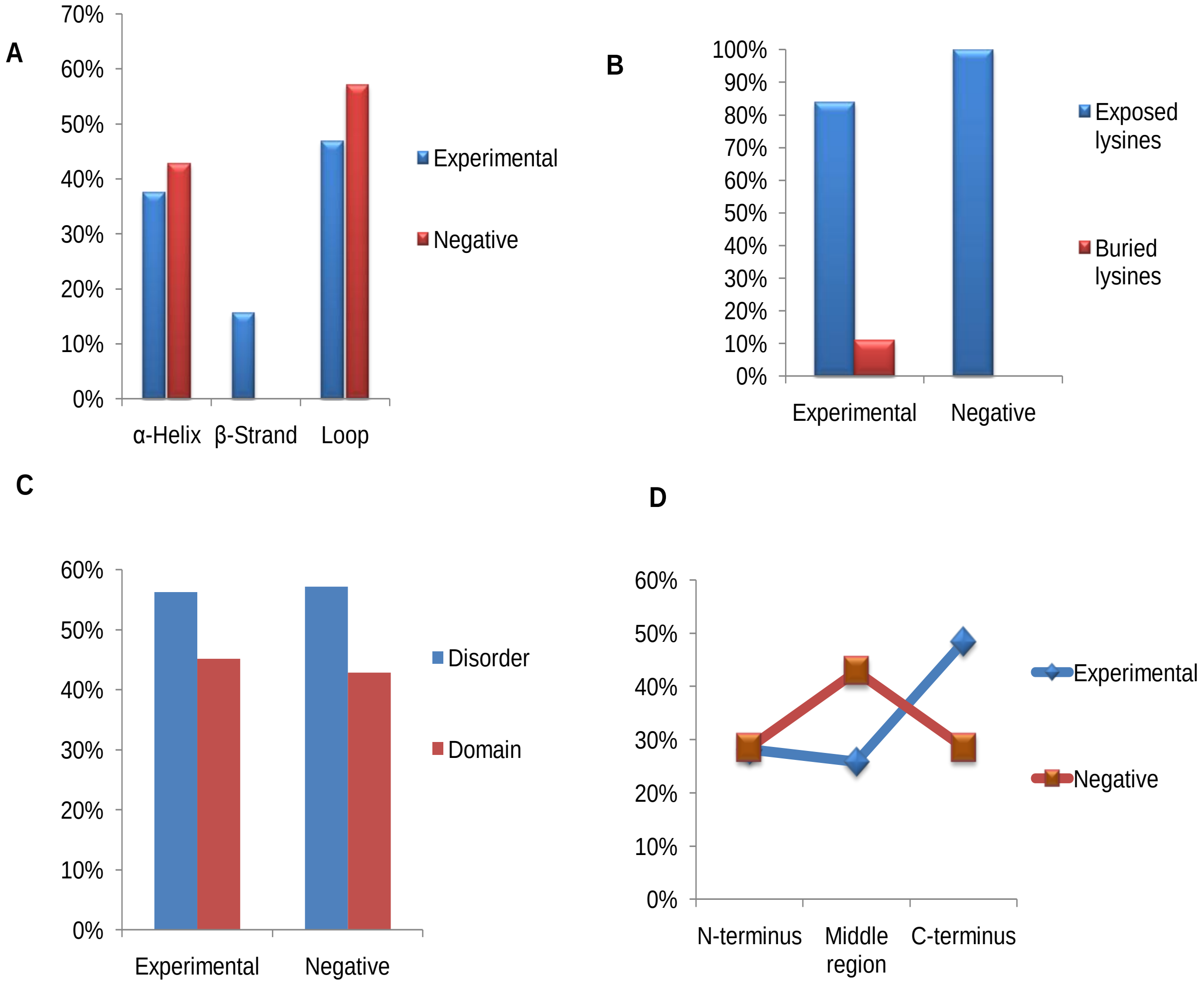

A
70%
60%
50%
40%
30%
20%
10%
0%
α-Helix
β-Strand
Loop
Experimental
Negative
B
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Experimental
Negative
Exposed lysines
Buried lysines
C
60%
50%
40%
30%
20%
10%
0%
Experimental
Negative
Disorder
Domain
D
60%
50%
40%
30%
20%
10%
0%
N-terminus
Middle region
C-terminus
Experimental
Negative

**Figure 3.** Structural context of predicted ubiquitination sites. **A.** Distribution based on secondary structure. **B.** Distribution based on solvent accessibility. **C.** Percentage distribution of predicted sites in disordered region and domain structure of protein. **D.** Percentage distribution of relative position of predicted ubiquitination site within the protein.

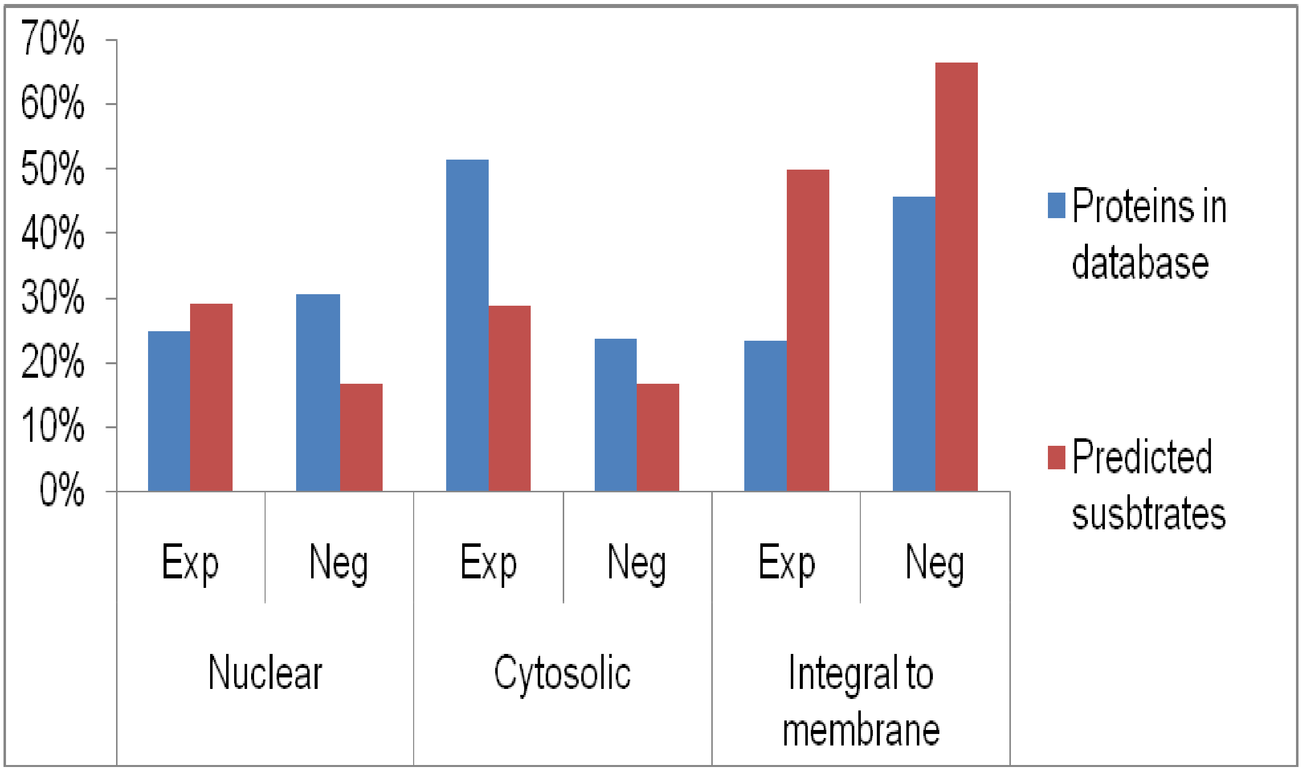

70%
60%
50%
40%
30%
20%
10%
0%
Exp
Neg
Exp
Neg
Exp
Neg
Nuclear
Cytosolic
Integral to membrane
Proteins in database
Predicted susbtrates

**Figure 4.** Sub-cellular localization of predicted TRAF6/p62 substrates in the database as compared to the proteins in the database.

# REFERENCES

1. Adamczak R, Porollo A, Meller J (2004) Accurate prediction of solvent accessibility using neural networks-based regression. Proteins. 56:753-767.

2. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33: W36–W38.

3. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, Dolinski K, Tyers M (2008) The BioGRID Interaction Database: 2008 update. Nucleic Acids Res (Database issue) D637-640.

4. Bueno MT, Garcia-Rivera JA, Kugelman JR, Morales E, Rosas-Acosta G, Llano M (2010) SUMOylation of the Lens-Epithelium-Derived Growth Factor/p75 Attenuates Its Transcriptional Activity on the Heat Shock Protein 27 Promoter. J Mol Biol 399: 221-239.

5. Catic A, Collins C, Church GM, Ploegh HL (2004) Preferred in vivo ubiquitination sites. Bioinformatics 20: 3302-3307.

6. Chew YC, Camporeale G, Kothapalli N, Sarath G, Zempleni J (2006) Lysine residues in N-terminal and C-terminal regions of human histone H2A are targets for biotinylation by biotinidase. J Nutr Biochem 17: 225-233.

7. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. Genome Research 14: 1188-1190.

8. Deshaies R, and Joazeiro C (2009) RING Domain E3 Ubiquitin Ligases. Annual Review of Biochemistry 78: 399-434.

9. Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. Bioinformatics. 23: 950-956.

10. Geetha T, Jiang J, Wooten MW (2005) Lysine 63 polyubiquitination of the nerve growth factor receptor TrkA directs internalization and signaling. Mol Cell 20: 301-312.

11. Geetha T, Kenchappa RS, Wooten MW, Carter BD (2005) TRAF6-mediated ubiquitination regulates nuclear translocation of NRIF, the p75 receptor interactor. EMBO J 24: 3859-3868.

12. Geetha T, and Wooten, MW (2002) Structure and functional properties of the ubiquitin binding protein p62. FEBS Lett 512: 19-24.

13. Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 8: R250.

14. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R; Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258-261.

15. Ishitani T, Takaesu G, Ninomiya-Tsuji J, Shibuya H, Gaynor RB, Matsumoto K (2003) Role of the TAB2-related protein TAB3 in IL-1 and TNF signaling. EMBO J 22: 6277–6288.

16. Jadhav T, Geetha T, Jiang J, Wooten MW (2008) Identification of a consensus site for TRAF6/p62 polyubiquitination. Biochem Biophys Res Commun 371: 521-524.

17. Jadhav T, and Wooten, MW (2009) Defining an Embedded Code for Protein Ubiquitination. J. Proteomics Bioinform 2: 316-333.

18. Jeon HB, Choi ES, Yoon JH, Hwang JH, Chang JW, Lee EK, Choi HW, Park ZY, Yoo YJ (2007) A proteomics approach to identify the ubiquitinated proteins in mouse heart. Biochem Biophys Res Commun 357: 731-736.

19. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195-202.

20. Lamothe B, Besse A, Campos AD, Webster WK, Wu H, Darnay BG (2007) Site-specific Lys-63-linked tumor necrosis factor receptor-associated factor 6 auto-ubiquitination is a critical determinant of IkappaB kinase activation. J Biol Chem 282: 4102–4112.

21. Lieutaud P, Canard B, Longhi S (2008) MeDor: a metaserver for predicting protein disorder. BMC Genomics 9: S25.

22. Li W, Bengtson MH, Ulbrich A, Matsuda A, Reddy VA, Orth A, Chanda SK, Batalov S, Joazeiro CA (2008) Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling. PLoS One. 3:1487.

23. Lin YS, Hsu WL, Hwang JK, Li WH (2007) Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol Biol Evol 24: 1005-1011.

24. Liu YC (2004) Ubiquitin ligases and the immune response. Annu Rev Immunol. 22: 81-127.

25. Loots G, Ovcharenko I (2007) ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. Bioinformatics. 23 :122-124.

26. Maathuis FJ (2008) Conservation of protein phosphorylation sites within gene families and across species. Plant Signal Behav 3: 1011-1013.

27. Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res., 33 (Database issue) D54-58.

28. Matsumoto M, Hatakeyama S, Oyamada K, Oda Y, Nishimura T, Nakayama KI (2005) Large-scale analysis of the human ubiquitin-related proteome. Proteomics 5: 4145- 4151.

29. Morén A, Hellman U, Inada Y, Imamura T, Heldin CH, Moustakas A (2003) Differential ubiquitination defines the functional status of the tumor suppressor Smad4. J Biol Chem 278: 33571-33582.

30. Ning S, Campos AD, Darnay BG, Bentz GL, Pagano JS (2008) TRAF6 and the three C-terminal Lysine sites on IRF7 are required for its ubiquitination-mediated activation by the tumor necrosis factor receptor family member latent membrane protein 1. Mol Cell Biol 28: 6536-6546.

31. Pang CN, Hayen A, Wilkins MR (2007) Surface accessibility of protein post-translational modifications. J Proteome Res 6: 1833-1845.

32. Pearson ES, and Hartley HD (1972) 'Biometrika' tables for statisticians 1. London: Cambridge University Press.

33. Pearson ES, and Hartley HD (1972) 'Biometrika' tables for statisticians 2. London: Cambridge University Press.

34. Peng J, Schwartz D, Elias JE, Thoreen CC, Cheng D, Marsischky G, Roelofs J, Finley D, Gygi SP (2003) A proteomics approach to understanding protein ubiquitination. Nat Biotechnol 21: 921-926.

35. Schneider TD, and Stephens RM (1990) Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res 18: 6097-6100.

36. Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. PNAS 95: 5857-5864.

37. Wu S, Liu T, Altman RB (2010) Identification of recurring protein structure microenvironments and discovery of novel functional sites around CYS residues. BMC Struct Biol 10: 4-22.

38. Willis MS, Schisler JC, Patterson C (2008) Appetite for destruction: E3 ubiquitin-ligase protection in cardiac disease. Future Cardiol 4: 65-75.

39. Zhou F, Xue Y, Lu H, Chen G, Yao X (2005) A genome-wide analysis of sumoylation-related biological processes and functions in human nucleus. FEBS Lett. 579: 3369-3375.

# CHAPTER 4. SUMMARY AND FUTURE DIRECTIONS

## SUMMARY

Most proteins in cells undergo post-translational modifications giving them structural and functional diversity to play important diverse roles in biological processes. Experimental identification and validation of posttranslational modifications (PTMs) is labor-intensive task and can be expensive in the absence of prior knowledge concerning PTMs. Analyzing 'ubiquitome' is one of the most exciting and challenging tasks in current proteomics research. The lack of curated datasets of ubiquitinated proteins presents the ultimate limiting factor in studying substrate selection mechanism in ubiquitination making it difficult to evaluate, and compare target sites. As more and more ligases are identified there exists an urgent need to rapidly and precisely identify enzyme-specific substrates to decode their selectivity and specificity (Li et al., 2008). Computational prediction of PTM sites has provided researchers with information on the high probability PTM sites for further experimental characterizations like PHOSIDA and NetPhos for phosphorylation (Gnad et al., 2007 and Blom et al., 2004), SUMOsp for SUMOylation (Xue et al., 2006) and NetAcet for prediction of N-acetyltransferase A substrates (Kiemer et al., 2005). Number of existing prediction tools for PTM sites were developed through various approaches using experimentally verified PTM sites and putative non-PTM sites as training datasets.

In this study, a computational tool was developed to predict Lysine ubiquitination sites from sequences using MATLAB programs and online based prediction softwares.  As more validated ubiquitinated sites from experimental data become available, and appropriate changes are made based on the available site data, reliable predictions can be made. The inclusion of structural information to improve the prediction tools could be another way to enhance the prediction performance as ubiquitination is an enzymatic process, and the interactions between target sites and enzymes concerned should be structurally satisfied. The model that I propose here can be applied to other E3 Ub ligases that are known to employ scaffold proteins to aid in their substrate selection process. One such example is DYRK2–EDVP E3 ligase complex where DYRK2 not only is it serves as adaptor for assembly of the active Ub ligase complex, but it also phosphorylates its substrate and primes the substrate for degradation (Maddika and Chen, 2009). Thus, use of bioinformatics methods to predict site modification *in silico* could yield more efficient results. These prediction tools should be closely integrated into the interpretation of proteomic experiments.

Here I identified the interactome of the active enzyme complex and studied the verified substrates for characterization of target sites to predict substrates. Fundamental understanding of their preferences for substrates would allow us to develop new research strategies to design drugs in context of various diseases they participate in.  As proteomics methods identify additional *in vivo* ubiquitination sites, prediction algorithms can be fine tuned and improved. A conserved motif that serves as a recognition determinant for TRAF6/p62 enzyme complex has been identified. Studies show some structural preferences for ubiquitination of the targeted proteins such as preferred choice of Lysines in loops and, and then for easily accessible Lysines in the α–helical region. This findings indicate a bias towards a consensus sequence motif for ubiquitination by a TRAF6/p62. Moreover, it appears that the active complex targets an accessible surface residue providing the selection process with a conformational recognition mechanism. The scaffold, p62, is important for recruiting substrates enabling the TRAF6 to scan for the easily accessible Lysine residues in the loops and helical structures on the surface of the substrate resulting in K63-polyubiquitination at a specific Lysine, if the flanking residues fit the consensus motif. The predicted Lysine 811 in TrkB was found to be ubiquitinated, and mutation of Lysine 811 diminished the formation of TRAF6/p62 complex that is necessary for effective ubiquitination. Downstream signaling was affected upon binding of BDNF to the mutant TrkB receptor. These findings reveal a possible selection process for targeting a specific Lysine residue by a single E3 ligase and underscore the role of the scaffold, p62, in this process. This report provides a strategy for studying how TRAF6 defines its Lysine specificity and reveals how scaffolds proteins, on which these complex chemical reactions take place, aid in selecting substrates. A total of 37 high probability TRAF6/p62 ubiquitination sites in 30 proteins were identified by this prediction approach. Structural analysis of these 30 predicted TRAF6/P62 substrates showed that the predicted ubiquitination sites were biased towards the C-terminal domain of the protein. Secondary structure analysis of the predicted sites revealed overall preference for loops and helices than beta-strands and solvent accessibility analysis of predicted Lysines revealed most of the predicted sites were exposed on the surface of the protein rather than being buried. There was high structural and phylogenetic conservation of predicted sites. Disordered regions inside as well as outside the domains of the proteins were preferred. This indicates that the high-confidence set of TRAF6/62 substrates and highly positive motif sites represent a reasonable site for modification by ubiquitin through TRAF6/p2 complex. Prediction of high amount of plasma membrane proteins in the high probability dataset indicates that the “code hypothesis” can be applied to other E3 ligases to predict their substrates.

This study links the classical approaches to find enzyme substrates through interacting proteins with modern computational approach. In conclusion, a holistic approach of using a combination of sequence motif data and structural determinants along with phylogenetic conservation can greatly aid in identifying the substrates and predicting putative ubiquitination sites. Lysine ubiquitination interplays actively with other post-translational modifications, either agonistically or antagonistically, to form a coded message for intramolecular signaling programs that are crucial for governing cellular functions. Given the intricacy of the ubiquitin system, research into its functions and mechanisms should continue to yield novel insights into cell regulation.

## FUTURE DIRECTIONS

Understanding the overall characteristics of motif specificity of TRAF6/p62 forms the foundation of bioinformatic computational approaches for identification of its substrates, and the functional characterization of these complex and the corresponding signal transduction pathways. Ubiquitination specificity is essential for the integrity of substrate recruitment and subsequent signal transduction events that strategically regulate other cellular processes. Understanding ubiquitination specificity will therefore contribute to understanding the roles of

E3 ligases in health and disease, and help identifying new therapeutic targets and strategies of E3 ligase inhibition and E3 ligase based drug development.

Results in this study indicate that the ubiquitination site prediction is closely correlated with the amino acid property around the ubiquitination site. And the computational tool developed in this work could be a powerful tool to investigate ubiquitination process preferences systemically. This approach makes it possible to find putative novel ubiquitination sites that have not (yet) been experimentally identified. Thus, in the absence of experimental data, the prediction of novel ubiquitination sites can be taken as the first method of an experimental design uncovering functionality of any protein of interest and elucidating its involvement in certain signaling cascades. Methods for computational prediction of peptide specificities and identification of substrates could be enhanced by combining different approaches and integrating various types of information. In addition, the prediction approach taken here combined with delicate experiments verifications will propel our understanding of the ubiquitination mechanisms.

Recent such tool developed, SLiMSearch, searches pre-defined SLiMs (Short Linear Motifs) in a protein sequence database taking into evolutionary relationships (Edwards RJ 2009). Therefore the next objective would be examining the search results in context to the ubiquitination linear motif described here and to compare the two approaches. Second objective will be to do a proteome wide search the for TRAF6/p62 ubiquitination sites. This search would lead us to only putative substrates and ubiquitination sites but also possibly putative interactors of either TRAF6 or p62. Thus also enrich our understanding of cellular interactome and proved insights into missing links within the cellular pathways and processes. Third objective would be to develop a convenient and comprehensive program, implement in an algorithm of Bayesian decision theory (BDT). The BDT approach has been extensively used to predict various PTMs prediction, such as of palmitoylation site (Xue et al., 2006) and PPSP prediction of PK-specific phosphorylation sites (Xue et al., 2006), and prediction of RNA structures (Ding, 2006).  Taken together, the prediction results lead us to fourth and final objective that would provide insightful and important for further experiments. This would be implemented by verifying proteins of interest from the high probability substrates and by study their biochemistry and signaling pathways. Thus combination of computational and experimental further objectives could propel our understanding of ubiquitination dynamics into a new phase.

## REFERENCES

1. Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, and Brunak S (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics. 4: 1633-1649.

2. Britton KL, Baker PJ, Fisher M, Ruzheinikov S, Gilmour DJ, Bonete MJ, Ferrer J, Pire C, Esclapez J, Rice DW (2006) Analysis of protein solvent interactions in glucose dehydrogenase from the extreme halophile Haloferax mediterranei. Proc Natl Acad Sci U S A. 103: 4846-4851.

3. RJ Edwards (2009) Last modified 04th Sep 2009.

4. Ding Y (2006) Statistical and Bayesian approaches to RNA secondary structure prediction. RNA. 12:323-331.

5. Kiemer L, Bendtsen JD, Blom N (2005) NetAcet: Prediction of N-terminal acetylation sites. Bioinformatics. 2005 21:1269-1270.

6. Li W, Bengtson MH, Ulbrich A, Matsuda A, Reddy VA, Orth A, Chanda SK, Batalov S, Joazeiro CA (2008) Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling. PLoS ONE 3: e1487.

7. Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 8: R250.

8. Maddika S, Chen J (2009) Protein kinase DYRK2 is a scaffold that facilitates assembly of an E3 ligase. Nat Cell Biol. 11: 379-381.

9. Xue Y, Zhou F, Fu C, Xu Y, Yao X (2006) SUMOsp: a web server for sumoylation site prediction.  Nucleic Acids Res 34: W254-257.

10. Yu Xue, Ao Li, Lirong Wang, Huanqing Feng, and Xuebiao Yao (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. BMC Bioinformatics. 20: 163.

11. Xue Y, Chen H, Jin C, Sun Z, Yao X (2006) NBA-Palm: prediction of palmitoylation site implemented in Naïve Bayes algorithm. BMC bioinformatics 7:458.

# APPENDIX

**Table T1.** List of proteins in the experimental dataset and the negative dataset in the database

| **Group I** | **Group II** | **Group III** | **Group IV** |
|---|---|---|---|
| Mapped TRAF6 substrates | Unmapped TRAF6 substrates | Traf6 interactors (TRAF6 substrates?!) | P62 interactors (TRAF6 substrates?!) |
| TrkA<br>TrkB<br>TrkC<br>NRIF | Tested:<br>TAU<br>UNC51.1(S/T_kinase)<br>IKBα<br>Hsp70<br>TRAF6<br>GLUR1 (AmpA)<br>APP | TRAK2<br>TRAKM (S/T_kinase)<br>MALT1<br>NIK<br>A20 (TrafB)<br>TAB2<br>TAB3<br>TRIP (TIR)<br>TRIF (CTD)<br>MAL (TIR)<br>Cezanne (TrafB) (DUB)<br>TRABID (TrafB) (DUB)<br>P62<br>ζPKC<br>IRAK1(TRAF) (S/T_kinase)<br>IRAK4<br>Pellino-1<br>Pellino-2<br>Pellino-3<br>TAK1<br>TAB1<br>RIP2<br>ZNF216 (ZnF-AN1)<br>TIZ<br>ACT1 (TRAF)(CTD)<br>MAST2<br>c-SRC<br>T6BP<br>ILPIPA<br>XIAP<br>UEV1A<br>Ubc13<br>USP7<br>SPOP<br>MUL<br>p75(NTR)<br>TTRAP(TRAF)<br>TIFA<br>SYK<br>TACI<br>TIRP<br>XEDAR<br>TROY<br>EDARADD<br>TRF7<br>ASK1<br>Spectrin<br>JUB<br>CYLD<br>KCNQ1<br>Proteasome subunits<br>PSMB5<br>PSMC1/S4<br>PSMC2/S7<br>PSMC3/TBP1<br>PSMD1/Rpn2<br>PSMD3/Rpn3<br>PSMD12<br>PSMD13<br>PSMC2/S5a | MAP2K5(S/T/Y_kinase)<br>PRKCI (S/T_kinase)<br>PRKCZ (S/T_kinase)<br>p56-LCK (Y_kinase)<br>RASA1<br>IRAK1(S/T_kinase)<br>ζPKC<br>NTRK2 (Y_kinase)<br>NTRK3 (Y_kinase)<br>PTPRJ (Y_phosphatase)<br>HCAP1<br>TRADD<br>TNFRSF1A<br>MAPKAPK5 (S/T_kinase)<br>IKBKB (S/T/Y_kinase)<br>Titin (S/T_kinase)<br>RIP(S/T_kinase)<br>NR2F2<br>TRAF6<br>PSMC2<br>JUB<br>LIMD1<br>TRIM55<br>GRB14<br>PAWR<br>NBR1 (PB1)<br>KV-BETA-2<br>ZIP1<br>ZIP2<br>ZIP3/p62<br>ρ1, ρ2, and ρ3 subunits of GABA$_{C}$ receptor 2<br>SNCA<br>ERCC5<br>ERCC2<br>ERCC3<br>MFN<br>P53<br>DRP1<br>KEAP<br>AKT |

**Table T1.** List of proteins in the experimental dataset and the negative dataset in the database (continued...)

| Group V | Negative dataset |
|---|---|
| FA fraction proteins from p62 knockout mice | |
| | DSCAM |
| | CD47 |
| | aSMase |
| 2',3'-cyclic-nucleotide 3'-phosphodiesterase I | BMPR2 |
| Actinin, alpha 1 | CA12 |
| akyrin 2 | CD79(Igbeta) |
| albumin (cow) | ErbB3 |
| ATP synthase beta-subunit (mouse) | EPHA8 |
| ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1 | EDA |
| beta-1-globin (mouse) | PTPRS |
| clathrin, heavy polypeptide (mouse) | SLC30A5 |
| Golli-mbp isoform 1 (mouse) | ADAM12 |
| golli-myelin basic protein precursor (mouse) | ADRA1B |
| Hemoglobin alpha, adult chain 1 (mouse) | RNF5 |
| hemoglobin beta minor chain (mouse) | ALK |
| heterogeneous nuclear ribonucleoprotein R (mouse) | MPL |
| Histone H4 (mouse) | IFNGR1 |
| Hnrpa3 protein (mouse) | CSF1R |
| Ina protein (mouse) | TACR2 |
| lamin A (mouse) | NOS2A |
| matrin3 (mouse) | LEPR |
| microtubule-associated protein 1B (human) | ADRBK1 |
| myosin H | BTK |
| myosin heavy chain 10, non-muscle (mouse) | AXL |
| myosin regulatory light polypeptide 9 | RPAIN |
| myosin, heavy polypeptide 10, non-muscle (mouse) | MKNK1 |
| neurofilament triplet M protein (mouse) | PTHLH |
| plectin isoform 1c (mouse) | ATF3 |
| ras GTPase-activating protein, synaptic (rat) | PTGG1 |
| Shc1_rat | HIF1A |
| similar to Spectrin alpha chain | MITF |
| spectrin alpha 2 (mouse) | CDC25C |
| Spectrin alpha chain | PCNA |
| spectrin beta 1 | FANCD2 |
| spectrin beta 2 isoform 1(mouse) | SMAD5 |
| spectrin beta 2 isoform 2 | EPB41 |
| spectrin beta 3 | UPFB3 |
| tubulin, beta 2 | BRCA1 |
| Tubulin, beta 2C (mouse) | Androgen Receptor |
| tubulin, beta 3 | RDM1 |
| tubulin, beta 3 | AIRE |
| similar to Tubulin, alpha 3c isoform 1 | ZNF677 |
| vesicle-fusing ATPase | ANG |
| H2afy protein | MTG16 |
| beta spectrin | NUMA1 |
| gamma-actin | NCL |
| | FUS |
| | KRT8 |
| | VIM |
| | CORO7 |
| | GOLGA2 |
| | ACO1 |
| | ST3GAL1 |

**Table T2.** Distribution of number of positive hits at seven variable positions in the consensus motif

| Frequency band of positive motif hits | # of hits in the Experimental dataset [in 157 proteins] | # of hits in the Negative dataset [in 54 proteins] |
|---|---|---|
| 1 | 1719 | 512 |
| 2 | 2381 | 713 |
| 3 | 1830 | 514 |
| 4 | 805 | 255 |
| 5 | 210 | 62 |
| 6 | 26 | 8 |
| 7 | 8 | 0 |

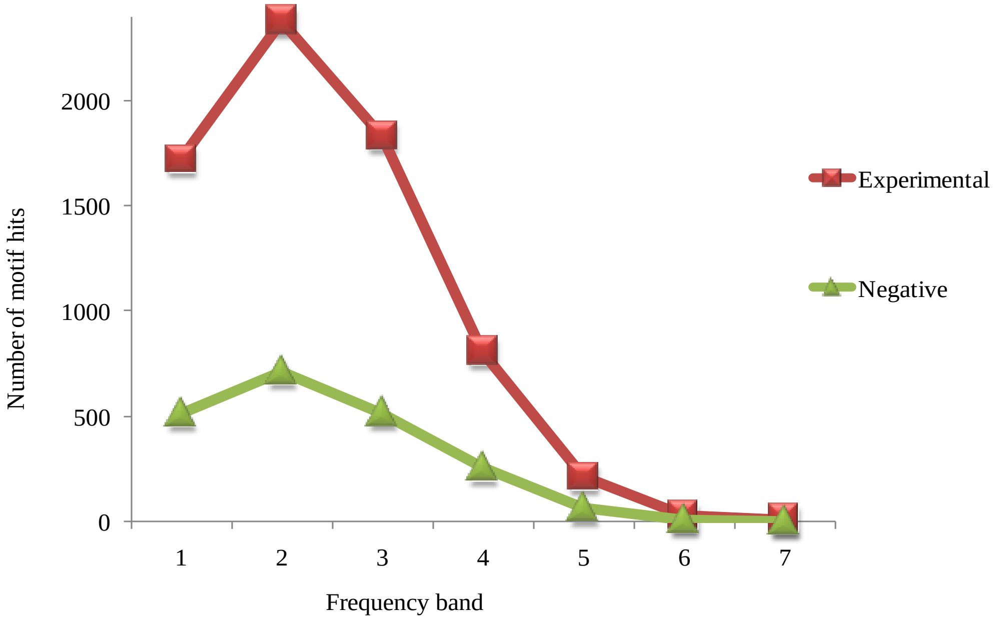


**Figure S1.** Frequency distribution curve of number of positive hits at seven variable positions in the consensus motif

**Table T3.** Localization of proteins in the Group I and Group II of the experimental dataset

| | | |
|---|---|---|
| **Group I** | | |
| | | |
| **Protein** | **Primary Localization** | **Secondary Localization** |
| TrkA | Plasma membrane | Cytoplasm, endosome |
| TrkB | Plasma membrane | Cytoplasm |
| TrkC | Plasma membrane | Cytoplasm |
| NRIF | Nucleus | Cytoplasm |
| | | |
| **Group II** | | |
| | | |
| **Protein** | **Primary Localization** | **Secondary Localization** |
| TAU | Cytoplasm | Plasma membrane, nucleus |
| UNC51.1 | Endoplasmic reticulum | Golgi apparatus, cytoplasm |
| IKBa | Cytoplasm | Nulcues, Mitochondrion |
| Hsp70 | Golgi apparatus, cytoplasm | Plasma membrane, nucleus, extracelllular |
| TRAF6 | Cytoplasm | Plasma membrane |
| GLUR1 | Plasma membrane | Cytoplasm |
| APP | Plasma membrane | Nulcues, Vesicle |

**Table T4.** Localization of proteins in the Group III of the experimental dataset

| Group III (TRAF6 interactors) | Primary localization | Secondary Localization |
|---|---|---|
| SYK | Plasma membrane | Cytoplasm |
| TACI | Plasma membrane | |
| TIRP | Plasma membrane | Cytoplasm, Golgi body |
| XEDAR | Plasma membrane | |
| TROY | Plasma membrane | |
| MAL | Plasma membrane | Cytoplasm, Endoplasmic Reticulum, Golgi body, Mitochondria, Endosome |
| IRAK1 | Plasma membrane | Cytoplasm |
| IRAK4 | Plasma membrane | Cytoplasm |
| MAST2 | Plasma membrane | Cytoplasm, Cytoskeleton |
| Tak1 | Plasma membrane | Nucleus |
| MALT1 | Nucleus | |
| NIK | Nucleus | Cytoplasm |
| A20 | Nucleus | Cytoplasm |
| Cezanne | Nucleus | Cytoplasm |
| TRABID | Nucleus | Cytoplasm |
| ZNF216 | Nucleus | |
| TIZ | Nucleus | Cytoplasm |
| T6BP | Nucleus | |
| ILPIPA | Nucleus | Cytoplasm |
| XIAP | Nucleus | Cytoplasm |
| USP7 | Nucleus | |
| SPOP | Nucleus | |
| TTRAP | Nucleus | |
| IRF7 | Nucleus | Cytoplasm |
| Pellino 1 | Cytoplsam | |
| Pellino 2 | Cytoplsam | |
| Pellino 3 | Cytoplsam | |
| TAB1 | Cytoplsam | |
| RIP2 | Cytoplsam | |
| MUL | Cytoplsam | Peroxisome |
| EDARADD | Cytoplsam | |
| TRIP | Cytoplsam | |
| TRIF | Cytoplsam | |
| ASK1 | Cytoplsam | |
| ACT1 | Cytoplsam | |
| JUB | Centrosome | Cytoplasm |
| TRAKM | Cytoplsam | |
| TAB2 | Cytoplsam | |
| TAB3 | Cytoplsam | |
| P62 | Cytoplsam | Nucleus, Late endosome |
| zPKC | Cytoplsam | |
| c-SRC | Cytoplsam | |
| UEV1A | Nucleus | |
| Ubc13 | Cytoplsam | |
| p75(NTR) | Plasma membrane | |
| TIFA | Plasma membrane | |
| Spectrin | Cytoplsam | |

**Table T5.** Localization of proteins in the Group IV of the experimental dataset

| Group IV (P62 interactors) | Primary localization | Secondary Localization |
|---|---|---|
| p56-LCK | Plasma membrane | Cytoplasm |
| PTPRJ | Plasma membrane | |
| TNFRSF1A/TNFR1 | Plasma membrane | Golgi body |
| MAP2K5 | Plasma membrane | Nucleus, Cytoplasm |
| RASA1 | Plasma membrane | Nucleus |
| IKBKB | Plasma membrane | Cytoplasm |
| JUB | Plasma membrane | Nucleus, Cytoplasm, Centrosome |
| GRB14 | Plasma membrane | Cytoplasm, ER, Golgi body, Endosome |
| NTRK1 | Plasma membrane | |
| NTRK2 | Plasma membrane | |
| NTRK3 | Plasma membrane | |
| alpha subunit of GABA Receptor2 | Plasma membrane | |
| Beata subunit of GABA Receptor2 | Plasma membrane | |
| Gamma subunits of GABA receptor 2 | Plasma membrane | |
| NR2F2 | Plasma membrane | Nucleus |
| KV-BETA-2 | Plasma membrane | Cytoplasm |
| PRKCI | Plasma membrane | Nucleus, Cytoplasm |
| PRKCZ | Plasma membrane | Nucleus, Cytoplasm, Late endosome and Microsome |
| IRAK1 | Cytoplasm | Nucleus, Cytoplasm |
| zPKC | Cytoplasm | |
| HCAP1 | Nucleus | Nucleus, Cytoplasm, Nucleolus |
| MAPKAPK5/p38 kinase | Cytoplasm | Nucleus, Cytoplasm |
| Titin | Cytoplasm | Nucleus, Cytoplasm |
| LIMD1 | Cytoplasm | Nucleus, Cytoplasm |
| TRIM55 | Cytoplasm | Nucleus, Cytoplasm |
| PAWR | Cytoplasm | Nucleus, Cytoplasm |
| SNCA | Cytoplasm | Nucleus, Cytoplasm |
| ZIP3/p62 | Cytoplasm | Nucleus, Cytoplasm, Late endosome |
| ZIP1 | Cytoplasm | |
| ZIP2 | Cytoplasm | |
| RIPK1/RIP | Cytoplasm | |
| TRAF6 | Cytoplasm | |
| PSMC2 | Cytoplasm | |
| NBR1 | Cytoplasm | |
| TRADD | Cytoplasm | |

**Table T6.** Localization of proteins in the Group V (FA fraction proteins from p62 KO mice) of the experimental dataset

| Group V (Protein Name) | Primary localization | Secondary localization |
|---|---|---|
| 2',3'-cyclic-nucleotide 3'-phosphodiesterase I | Cytoplasm, Extracellular space | Plasma membrane |
| Actinin, alpha 1 | Cytoplasm | Mitochondrial membrane, Mitochondria, Nucleus, Cytoskeleton |
| akyrin 2 | Extracellular | Cytoplasm |
| albumin (cow) | Extracellular | |
| ATP synthase beta-subunit (mouse) | Mitochondrion | |
| ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1 | Mitochondrion | Extracellular, Zymogen granule |
| beta-1-globin (mouse) | Extracellular | |
| clathrin, heavy polypeptide (mouse) | Clathrin-coated vesicle | |
| glial fibrillary acidic protein, astrocyte (mouse) | | |
| Golli-mbp isoform 1 (mouse) | Plasma membrane | Cytosol, Nucleus |
| golli-myelin basic protein precursor (mouse) | Cytoplasm | |
| Hemoglobin alpha, adult chain 1 (mouse) | Cytoplasm | |
| hemoglobin beta minor chain (mouse) | Cytoplasm | |
| heterogeneous nuclear ribonucleoprotein R (mouse) | Nucleus | Nucleous, Mitochondria |
| Histone H4 (mouse) | Nucleus | |
| Hnrpa3 protein (mouse) | Nucleus | Cytoplasm |
| Hsc70-ps1 (rat) | Cytoplasm | |
| Ina protein (mouse) | Cytoplasm | |
| Lamin A (mouse) | Nucleus | Cytoplasm, Nucleolus |
| Matrin3 (mouse) | Nucleus | Cytoplasm, Nucleolus |
| Microtubule-associated protein 1B (human) | Cytoplasm | Plasma membrane, Nucleus |
| mKIAA0788 protein (mouse) | Cytoplasm | |
| Myosin H | Cytoplasm | Cytoplasm |
| Myosin heavy chain 10, non-muscle (mouse) | Cytoplasm | Cytoskeleton |

**Table T6.** Localization of proteins in the Group V (FA fraction proteins from p62 KO mice) of the experimental dataset *(continued...)*

| | | |
|---|---|---|
| Myosin regulatory light polypeptide 9 | Cytoplasm | Cytoskeleton |
| Myosin, heavy polypeptide 10, non-muscle (mouse) | Cytoplasm | Cytoskeleton |
| Na+/K+ -ATPase alpha 3 subunit (mouse) | Cytoskeleton | |
| Neurofilament protein, high molecular weight subunit (NF-H) (mouse) | Cytoskeleton | |
| Neurofilament triplet M protein (mouse) | Cytoskeleton | |
| Neurofilament, heavy polypeptide (mouse) | Cytoskeleton | |
| Nonmuscle myosin heavy chain | Cytoskeleton | |
| PL10 protein (mouse) | Cytoskeleton | |
| Plectin isoform 1c (mouse) | Cytoskeleton | Cytoplasm, Nucleus, Nucleolus, Plasma Membrane, Mitochondrion |
| Ras GTPase-activating protein, synaptic (rat) | Cytoplasmic vesicle | |
| Shc1_rat | Cytoplasm | PM, Endoplasmic reticulum, |
| Similar to CG31613-PA (rat) | Cytoplasm | |
| Similar to Spectrin alpha chain | Cytoskeleton | |
| Spectrin alpha 2 (mouse) | Cytoskeleton | |
| Spectrin alpha chain | Cytoskeleton | |
| Spectrin beta 1 | Cytoskeleton | |
| Spectrin beta 2 isoform 1(mouse) | Cytoskeleton | |
| Spectrin beta 2 isoform 2 | Cytoskeleton | |
| Spectrin beta 3 | Cytoskeleton | |
| Tubulin, beta 2 | Cytoskeleton | |
| Tubulin, beta 2C (mouse) | Cytoskeleton | |
| Tubulin, beta 3 | Cytoskeleton | |
| Tubulin, beta 3 | Cytoskeleton | |
| Similar to Tubulin, alpha 3c isoform 1 | Cytoskeleton | |
| Vesicle-fusing ATPase | Cytoplasm | Golgi body, Plasma membrane, Cytoplasm |
| H2afy protein | Centrosome | Nucleus |
| Beta spectrin | Cytoskeleton | |
| Gamma-actin | Cytoskeleton | |

**Table T7.** Localization of proteins in the negative dataset

| **Protein Name** | **Primary localization** | **Secondary localization** |
|---|---|---|
| CD34 | Plasma membrane | Extracellular |
| DSCAM | Plasma membrane | Extracellular |
| CD47 | Plasma membrane, cell surface | |
| aSMase | lysosome | ER, Extracellular, Plasma membrane |
| BMPR2 | Plasma membrane | |
| CA12 | Plasma membrane | |
| CD79(Igbeta) | Plasma membrane | Cytoplasm |
| ErbB3 | Extracellular | Plasma membrane |
| EPHA8 | Plasma membrane | |
| EDA | Plasma membrane | Cytoskeleton, Extracellular |
| PTPRS | Plasma membrane | |
| SLC30A5 | Plasma membrane | Golgi apparatus, Secretory body |
| ADAM12 | Plasma membrane | |
| ADRA1B | Plasma membrane | |
| RNF5 | Plasma membrane | Nucleus, Endoplasmic Reticulum membrane |
| ALK | Plasma membrane | Cell surface |
| MPL | Plasma membrane | |
| IFNGR1 | Plasma membrane | |
| CSF1R | Plasma membrane | |
| TACR2 | Plasma membrane | |
| NOS2A | Cytoplasm | Plasma membrane |
| LEPR | Plasma membrane | Cell surface, Early endosome |
| ADRBK1 | Cytoplasm | Plasma membrane |
| BTK | Cytoplasm | Plasma membrane, nucleus |

**Table T7.** Localization of proteins in the negative dataset *(continued...)*

| | | |
|---|---|---|
| AXL | Plasma membrane | Extracellular |
| RPAIN | Cytoplasm | Nucleus |
| MKNK1 | Cytoplasm | Nucleus |
| PTHLH | Extracellular | Nucleus, Cytoplasm, Nucleolus |
| ATF3 | Nucleus | |
| PTTG1 | Cytoplasm | Nucleus |
| HIF1A | Nucleus | Nucleolus, Cytoplasm |
| MITF | Nucleus | Cytoplasm |
| CDC25C | Nucleus | Cytoplasm |
| PCNA | Nucleus | Cytoplasm, Nucleolus |
| FANCD2 | Nucleus | Mitochondrion |
| SMAD5 | Nucleus | Cytoplasm, Nucleolus |
| EPB41 | Nucleus | Cytoplasm, Plasma membrane, Centrosome |
| UPF3B | Nucleus | Cytoplasm, Nucleolus |
| BRCA1 | Nucleus | Cytoplasm, Mitochondrion, Centrosome, Perinuclear region |
| Androgen Receptor | Nucleus | Cytoplasm, Membrane-associated |
| RDM1 | Nucleus | |
| AIRE | Nucleus | Cytoplasm |
| ZNF677 | Nucleus | |
| ANG | Extracellular | Nucleolus, Nucleus |
| MTG16 | Nucleus | Golgi apparatus, Cytoplasm, Nucleolus |
| NUMA1 | Nucleus | Nucleolus, Cytoplasm, Mitochondrion, Microtubule |
| NCL | Nucleolus | Nucleus, Cytplasm, Plasma membrane |
| FUS | nucleus | Cytoplasm, Nucleolus, Mitochondrion |
| KRT8 | Cytoplasm | Nucleolus, Extracellular, Cytoskeleton, Nucleus |
| VIM | Cytoskeleton | Intermediate filament, Nucleolus, Nucleus, Cytoplasm, Membrane fraction, Extracellular, ER, Golgi body |
| CORO7 | Golgi membrane | Cytoplasm |
| GOLGA2 | Golgi membrane | |
| ACO1 | Cytoplasm | Golgi membrane, Endoplasmic Reticulum |
| ST3GAL1 | Golgi membrane | |

**Table T8.** List of proteins with positive hits at 5 or more variable positions in the experimental dataset

| Protein | Match | Sequence match | Lysine Position |
|---|---|---|---|
| TrkA | 5 | akllaggedv | 612 |
| | 7 | gkgsglqghi | 485 |
| TrkB | 6 | vkfygvcveg | 601 |
| | 7 | akaspvyldi | 811 |
| TrkC | 5 | mkgpvavisg | 465 |
| | 7 | vkfygvcgdg | 602 |
| | 7 | gkatpiyldi | 815 |
| Nrif | 7 | vkfedvsltf | 19 |
| | 5 | gkafrqsshl | 779 |
| Tau | 5 | akgqdaplef | 293 |
| | 5 | vkgdlaflnf | 98 |
| Hsp70 | 6 | akaaaigidl | 3 |
| | 5 | vkatagdthl | 220 |
| | 6 | akldkaqihd | 325 |
| | 5 | gkankititn | 497 |
| Traf6 | 5 | akreilslmv | 124 |
| | 5 | akmetqsmyv | 319 |
| | 5 | wkignfgmhl | 365 |
| AMPA | 5 | fkesganvtg | 244 |
| | 5 | dkgecgsggg | 784 |
| RASA1 | 5 | lkgdmfivhn | 303 |
| PTPRJ | 5 | ikavsisptn | 126 |
| | 5 | dkaitlqgli | 589 |
| | 5 | ikayaviltt | 848 |
| PSMC2 | 5 | fkiharsmsv | 356 |
| PRKCZ | 5 | rklyranghl | 124 |
| | 6 | lkldnvllda | 378 |
| p56LCK | 5 | lkqgsmspda | 276 |
| NTRK3_human | 5 | mkgpvavisg | 465 |
| | 7 | vkfygvcgdg | 602 |
| | 7 | gkatpiyldi | 829 |
| NTRK2_human | 5 | gkvksrqgvg | 474 |
| | 6 | vkfygvcveg | 618 |
| | 7 | akaspvyldi | 828 |
| NR2F2 | 6 | lkfmwgnltl | 413 |
| MAPKAPK5 | 5 | rkimtgsfef | 257 |
| MAP2K5 | 5 | gkilavkvil | 190 |
| | 6 | vkvillditl | 195 |
| KVBETA2 | 5 | gkaevvlgni | 94 |
| | 6 | aklkelqaia | 288 |

**Table T8.** List of proteins with positive hits at 5 or more variable positions in the experimental dataset *(continued..)*

| | | | |
|---|---|---|---|
| IKBKB | 5 | lkariqqdtg | 337 |
| GABRR1 | 6 | vkavdvymwv | 335 |
| GABRR2 | 6 | vkavdiylwv | 322 |
| GABRR1 | 6 | ikavdiylwv | 336 |
| ERRC5 | 5 | gkilavdisi | 25 |
| | 6 | skmhgmsfdv | 313 |
| | 6 | gkgipftatl | 438 |
| | 5 | kklrtlqltp | 917 |
| | 5 | gkekmvlvta | 1157 |
| ERRC5 | 5 | akdyrlqmpl | 59 |
| ERRC3 | 5 | akmfrrvlti | 449 |
| | 5 | skvgdtsfdl | 609 |
| TRIM55 | 5 | ekfdylygil | 214 |
| AKT | 5 | gkgtfgkvil | 158 |
| | 5 | lklenlmldk | 276 |
| TNFRSFA1 | 5 | vkgtedsgtt | 203 |
| PRKC1 | 5 | lkldnvllds | 371 |
| P53 | 5 | aktcpvqlwv | 139 |
| MFN1 | 6 | wkllsvsltm | 613 |
| A20 | 6 | lkvggiylpl | 228 |
| ASK1 | 5 | gkldfgettv | 134 |
| | 5 | akaldimipm | 370 |
| | 5 | gkgtygivya | 688 |
| | 5 | ikifmeqvpg | 751 |
| | 5 | dkgprgygka | 853 |
| | 5 | fkvgmfkvhp | 893 |
| | 5 | lkvdpfsfkt | 992 |
| CEZANNE | 5 | vkwiplssda | 432 |
| CYLD | 5 | lkvpkgsigq | 40 |
| | 5 | akgkknqigl | 64 |
| | 5 | gkeslgyfvg | 258 |
| | 5 | gkkkgiqghy | 590 |
| | 5 | gkikqfcktc | 812 |
| ILPIPA | 5 | ikashilisg | 186 |
| IRAK4 | 5 | vkklaamvdi | 213 |
| KCNQ1 | 5 | akkcpfslel | 32 |
| MALT1 | 5 | gkpliakldm | 709 |
| MAST2 | 5 | skiglmsltt | 658 |
| NIK | 5 | gkmarvcwkg | 128 |
| | 5 | vkvqiqslng | 862 |

**Table T8.** List of proteins with positive hits at 5 or more variable positions in the experimental dataset *(continued..)*

| | | | |
|---|---|---|---|
| PEL1 | 5 | wktsdgqmdg | 174 |
| PSMB5 | 5 | fkfrhgviva | 66 |
| PSMC2 | 5 | fkiharsmsv | 356 |
| PSMD12 | 5 | lksvvlyvil | 268 |
| | 5 | akvdrlagii | 405 |
| PSMD13 | 5 | lklnigdlqv | 122 |
| PSMD1 | 5 | ikilsgemai | 327 |
| | 5 | akfgailaqg | 727 |
| PSMD4 | 5 | lkkekvnvdi | 132 |
| SPOP | 5 | fkfsilnakg | 103 |
| SRC | 5 | vklgqgcfge | 275 |
| | 5 | eklvqlyavv | 324 |
| SYK | 5 | mkgsevtaml | 577 |
| TAB1 | 5 | ykvkygytdi | 247 |
| TAB2 | 5 | rklsmgsdda | 522 |
| TAB3 | 5 | fkitvgratt | 456 |
| | 5 | rkarrisvts | 640 |
| TACI | 5 | lklsadqval | 154 |
| TRAK2 | 5 | vkplegsqtl | 562 |
| TRIP | 6 | gkaemlcstl | 127 |
| | 5 | kkltmlqetl | 270 |
| UBC13 | 6 | dklgricldi | 82 |
| UVE1A | 6 | mkgtcvegti | 312 |
| XIAP | 5 | eklckicmdr | 448 |
| ZNF675 | 5 | kafnqsshl | 263 |
| | 5 | gkaftqsstl | 319 |
| 2',3'-cyclic-nucleotide 3'-phosphodiesterase I | 6 | gkafklsisa | 259 |
| | 6 | gkgkpvpihg | 379 |
| Actinin, alpha 1 | 5 | fkaclislgy | 772 |
| Akyrin 2 | 5 | gkvrlpalhi | 186 |
| | 5 | gkteivqlll | 505 |
| Albumin (cow) | 5 | eklftfhadi | 528 |
| ATP synthase beta-subunit (mouse) | 5 | ikipvgpetl | 133 |
| | 5 | akggkiglfg | 198 |
| | 5 | gkiglfggag | 201 |
| | 5 | akahggysvf | 225 |
| | 5 | skvalvygqm | 265 |
| | 5 | gklvplketi | 480 |

**Table T8.** List of proteins with positive hits at 5 or more variable positions in the experimental dataset *(continued..)*

| | | | |
|---|---|---|---|
| ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1 | 5 | vkrtgaivdv | 132 |
| | 5 | gkgpigsktr | 161 |
| | 5 | kklyciyvai | 241 |
| Beta-1-globin (mouse) | 5 | lkgtfaslse | 82 |
| | 5 | qkvmagvata | 132 |
| Clathrin, heavy polypeptide (mouse) | 5 | mkahtmtddv | 100 |
| | 5 | akqkwllltg | 161 |
| | 5 | rkgqvlsvcv | 321 |
| | 5 | ykaiqfylef | 1406 |
| Golli-myelin basic protein precursor (mouse) | 5 | pkipsisthi | 426 |
| Hemoglobin alpha, adult chain 1 (mouse) | 5 | ikaawgkigg | 12 |
| | 5 | fkllshcllv | 100 |
| Hemoglobin beta minor chain (mouse) | 5 | lkgtfaslse | 82 |
| | 5 | qkvvagvata | 132 |
| Heterogeneous nuclear ribonucleoprotein R (mouse) | 5 | gkhlgvcisv | 235 |
| | 5 | skvteglvdv | 268 |
| | 5 | vkvwgnvvtv | 315 |
| Histone H4 (mouse) | 5 | gkggkglgkg | 6 |
| Hsc70-ps1 (rat) | 5 | skgpavgidl | 3 |
| | 5 | akldksqihd | 325 |
| Ina protein (mouse) | 5 | lkaqqrdvdg | 197 |
| Lamin A (mouse) | 5 | akleaalgea | 171 |
| Microtubule-associated protein 1B (human) | 5 | iklnsasilp | 213 |
| | 5 | gkaaeavaaa | 790 |
| | 5 | lkaeevdvtk | 844 |
| mKIAA0788 protein (mouse) | 5 | gkhinmdgti | 296 |
| | 5 | lklegfalma | 844 |
| | 5 | gktihkyvhl | 945 |

**Table T8.** List of proteins with positive hits at 5 or more variable positions in the experimental dataset *(continued..)*

| | | | |
|---|---|---|---|
| Myosin H | 5 | dklraaciri | 764 |
| | 5 | dkgeiaqayi | 1304 |
| | 5 | lkprgvavhl | 1493 |
| | 5 | vkvlnlytpv | 1778 |
| Myosin heavy chain 10, non-muscle (mouse) | 6 | gkfirinfdv | 244 |
| | 5 | lkitdiiiff | 785 |
| | 5 | lkdleaqiea | 1627 |
| | 5 | aklqelegav | 1800 |
| Na+/K+ -ATPase alpha 3 subunit (mouse) | 5 | ckvdnssltg | 202 |
| | 5 | ikvimvtgdh | 602 |
| | 5 | akacvihgtd | 651 |
| Neurofilament protein, high molecular weight subunit (NF-H) (mouse) | 5 | pkipsisthi | 425 |
| Neurofilament, heavy polypeptide (mouse) | 5 | lpkipsisthi | 426 |
| Nonmuscle myosin heavy chain | 6 | gkfirinfdv | 244 |
| | 5 | lkitdiiiff | 785 |
| | 5 | lkdlegqiea | 1627 |
| | 5 | aklqelegsv | 1800 |
| PL10 protein (mouse) | 5 | gkspilvata | 490 |
| | 5 | hklqnvqial | 128 |
| | 5 | lkippgyhpl | 385 |
| | 5 | kkikeiqntg | 697 |
| Plectin isoform 1c (mouse) | 5 | lkentayfqf | 745 |
| | 5 | lkdirlqlea | 1016 |
| | 5 | eklktislvi | 1113 |
| | 5 | lkklraqaea | 1159 |
| | 5 | gkfqgrtvti | 2941 |
| | 5 | ekiikivitv | 2979 |
| | 5 | ekvikiviti | 3308 |
| | 5 | lkkgllsaev | 3736 |
| | 5 | vkgerltvde | 3763 |

**Table T8.** List of proteins with positive hits at 5 or more variable positions in the experimental dataset *(continued..)*

| | | | |
|---|---|---|---|
| Ras GTPase-activating protein, synaptic (rat) | 5 | ggkgkggcpav | 377 |
| | 5 | gkeevasalv | 429 |
| | 5 | gkakdflsdm | 445 |
| Shc1_rat | 6 | lkfagmpitl | 116 |
| Similar to CG31613-PA (rat) | 5 | gkggkglgkg | 143 |
| | 5 | afkrafiyvds | 427 |
| Similar to Spectrin alpha chain | 5 | ikllqaqklv | 144 |
| | 5 | kkfeefqtdl | 190 |
| | 5 | lkglalqrqg | 241 |
| | 5 | dkvkalcaea | 310 |
| | 5 | vkalcaeadr | 312 |
| Spectrin alpha 2 (mouse) | 5 | ikllqaqklv | 144 |
| | 5 | kkfeefqtdl | 190 |
| | 5 | lkglalqrqg | 241 |
| | 5 | dkvkalcaea | 310 |
| | 5 | vkalcaeadr | 312 |
| | 5 | kkfddfqkd | 1112 |
| | 5 | akldensafl | 1951 |
| | 5 | kklleaqshf | 2058 |
| | 5 | rkvedlfltf | 2068 |
| Spectrin alpha chain | 5 | ikllqaqklv | 144 |
| | 5 | kkfeefqtdl | 190 |
| | 5 | lkglalqrqg | 241 |
| | 5 | dkvkalcaea | 310 |
| | 5 | vkalcaeadr | 312 |
| | 5 | kkfddfqkdl | 1132 |
| | 5 | ekiaalqafa | 1500 |
| | 5 | akldensafl | 1971 |
| | 5 | kklleaqshf | 2078 |
| | 5 | rkvedlfltf | 2088 |
| Spectrin beta 1 | 5 | akakaeqlsa | 1369 |
| | 5 | akaeqlsaar | 1371 |

**Table T8.** List of proteins with positive hits at 5 or more variable positions in the experimental dataset *(continued..)*

| | | | |
|---|---|---|---|
| Spectrin beta 2 isoform 1 (mouse) | 5 | mkvlllsqdy | 548 |
| | 5 | aklsdlqkea | 1011 |
| | 5 | skvdklyagl | 1675 |
| | 5 | ikekllqlte | 1989 |
| Spectrin beta 2 isoform 2 | 5 | mkvlllsqdy | 535 |
| | 5 | aklsdlqkea | 998 |
| | 5 | hkaqqyyfda | 1580 |
| | 5 | skvdklyagl | 1662 |
| | 5 | ikekllqlte | 1976 |
| Spectrin beta 3 | 5 | mkgrlqsqdl | 551 |
| | 5 | ekmdwlqlvl | 2005 |
| Tubulin, beta 2 | 5 | fkriseqfta | 379 |
| Tubulin, beta 2C (mouse) | 5 | lfkriseqfta | 379 |
| Tubulin, beta 3 | 5 | fkriseqfta | 379 |
| Vesicle-fusing ATPase | 5 | akqcigtmti | 89 |
| | 5 | lkgepasgkr | 161 |
| | 5 | vkgillygpp | 254 |
| | 5 | ekaeslqvtr | 469 |
| | 5 | dkmigfseta | 572 |
| | 5 | vkgkkvwigi | 699 |
| H2afy protein | 6 | qklqvvqadi | 196 |
| Beta spectrin | 5 | mkvlllsqdy | 534 |
| | 5 | aklsdlqkea | 993 |
| | 5 | hkaqqyyfda | 1573 |
| | 5 | skvdklyagl | 1655 |
| | 5 | ikekllqlte | 1968 |
| Gamma-actin | 5 | eklcyvaldf | 208 |
| MBP | 7 | fkgvdaqgtl | 169 |

**Table T9.** List of proteins with positive hits at 5 or more variable positions in the negative dataset

| Protein | Match | Sequence match | Position |
|---|---|---|---|
| DSCAM | 5 | gkirsqdvhi | 110 |
| | 5 | lklsdvqkev | 560 |
| | 6 | vkaaaasasm | 1196 |
| | 5 | akapariltf | 1283 |
| BMPR2 | 5 | lklleligrg | 204 |
| | 5 | lkqvdmyalg | 402 |
| ErbB3 | 5 | lkmcepcggl | 318 |
| EPHA8 | 5 | lkidtiaade | 144 |
| | 5 | lkavttratv | 491 |
| | 5 | gklpepqfya | 603 |
| EDA | 5 | fklhprsgel | 285 |
| | 5 | vkmvhadisi | 363 |
| PTPRS | 5 | ektvdvyghv | 1613 |
| SLC30A5 | 5 | lklgtaffmv | 67 |
| ADAM12 | 6 | lkpdavcahg | 459 |
| RNF5 | 5 | ekvvplygrg | 75 |
| ALK | 5 | lkvmeghgev | 971 |
| | 5 | hkvicfcdhg | 1003 |
| MPL | 5 | ikamggsqpg | 140 |
| IFNGR1 | 5 | gkigppkldi | 126 |
| | 5 | ekskevciti | 230 |
| CSF1R | 5 | rkvmsisirl | 185 |
| | 5 | gkvveatafg | 595 |
| | 5 | vkmlkstaha | 616 |
| NOS2A | 5 | fkaacetfdv | 678 |
| LEPR | 5 | lkitsggvif | 214 |
| | 5 | aksksvslpv | 592 |
| BTK | 5 | fkkrlflltv | 26 |
| AXL | 5 | akgvttsrta | 211 |
| | 5 | lkqpadcldg | 769 |
| MKNK1 | 5 | eklqggsila | 126 |
| HIF1A | 5 | dkasvmrlti | 56 |
| | 5 | mkaqmncfyl | 85 |
| | 5 | lkaldgfvmv | 94 |

**Table T9.** List of proteins with positive hits at 5 or more variable positions in the negative dataset *(continued...)*

| | | | |
|---|---|---|---|
| CDC25C | 5 | gkflgdsanl | 52 |
| | 5 | vkkkyfsgqg | 242 |
| PCNA | 5 | tkatplsstv | 217 |
| FANCD2 | 6 | vkllkisgii | 50 |
| | 6 | ikfilhsvta | 283 |
| | 5 | lkvrqlvmdk | 261 |
| | 5 | vkgildyldn | 515 |
| SMAD5 | 5 | gkgvhlyyvg | 332 |
| UPFB3 | 5 | ikvhrfllqa | 269 |
| BRCA1 | 5 | lkltnapgsf | 701 |
| Androgen Receptor | 5 | ckavsvsmgl | 241 |
| | 6 | gkvkpiyfht | 911 |
| AIRE | 5 | akgaqgaapg | 259 |
| ZNF677 | 5 | gkafkqcshl | 410 |
| MTG16 | 5 | lkwsmvcllm | 120 |
| | 5 | akmeralaea | 526 |
| NUMA1 | 5 | fklrefashl | 326 |
| | 5 | gklsqleehl | 386 |
| | 5 | akllaerghf | 449 |
| | 5 | akleilqqql | 616 |
| | 5 | rkveelqacv | 651 |
| | 5 | lkvtkgslee | 712 |
| | 5 | qklkavqaqg | 1571 |
| | 5 | lkavqaqgge | 1573 |
| NCL | 5 | akagknqgdp | 6 |
| | 5 | akndlavvdv | 333 |
| FUS | 5 | akaaidwfdg | 348 |
| KRT8 | 6 | lkgqraslea | 325 |
| | 5 | aklseleaal | 352 |
| | 5 | gklvsessdv | 472 |
| CORO7 | 5 | vklwrlpgpg | 103 |
| | 6 | skfrhaqgtv | 472 |
| GOLGA2 | 5 | vkllelqelv | 869 |
| ACO1 | 5 | gkfveffgpg | 276 |
| ST3GAL1 | 5 | lkvltflvlf | 10 |

**Table T10.** Secondary structure analysis of proteins with positive hits at 6 or more variable positions in the experimental dataset

| Protein | Match | Sequence match | Position | Secondary Structure | Accessibility |
|---|---|---|---|---|---|
| TrkA | 7 | gkgsglqghi | 485 | loop | exposed |
| TrkB | 6 | vkfygvcveg | 601 | helix | exposed |
|  | 7 | akaspvyldi | 811 | loop | exposed |
| TrkC | 7 | vkfygvcgdg | 602 | b strand | buried |
|  | 7 | gkatpiyldi | 815 | loop | exposed |
| Nrif | 7 | vkfedvsltf | 19 | loop | exposed |
| Hsp70 | 6 | akaaaigidl | 3 | loop | exposed |
|  | 6 | akldkaqihd | 325 | helix | exposed |
| PRKCZ | 6 | lkldnvllda | 378 | loop | exposed |
| NTRK3 | 7 | vkfygvcgdg | 602 | b strand | exposed |
| NTRK3 | 7 | gkatpiyldi | 829 | loop | exposed |
| NTRK2 | 6 | vkfygvcveg | 618 | helix | exposed |
| NTRK2 | 7 | akaspvyldi | 828 | loop | exposed |
| NBR1 | 6 | lkfmwgnltl | 413 | b strand | buried |
| MAP2K5 | 6 | vkvillditl | 195 | helix | buried |
| KVBETA2 | 6 | aklkelqaia | 288 | helix | buried |
| GABRR2 | 6 | vkavdiylwv | 322 | helix | exposed |
| GABRR1 | 6 | ikavdiylwv | 336 | helix | exposed |
| GABRR3 | 6 | vkavdvymwv | 325 | helix | exposed |
| ERCC5 | 6 | skmhgmsfdv | 313 | loop | exposed |
|  | 6 | gkgipftatl | 438 | loop | exposed |
| MFN1 | 6 | wkllsvsltm | 613 | helix | exposed |
| A20 | 6 | lkvggiylpl | 228 | loop | exposed |
| TRIP | 6 | gkaemlcstl | 127 | helix | exposed |
| UBC13 | 6 | dklgricldi | 82 | helix | exposed |
| USP7 | 6 | mkgtcvegti | 312 | loop | exposed |
| 2',3'-cyclic-nucleotide 3'-phosphodiesterase I | 6 | gkafklsisa | 259 | loop | exposed |
|  | 6 | gkgkpvpihg | 379 | loop | exposed |
| Myosin heavy chain 10, non-muscle (mouse) | 6 | gkfirinfdv | 244 | b strand | buried |
| Shc1_rat | 6 | lkfagmpitl | 116 | loop | exposed |
| H2afy protein | 6 | qklqvvqadi | 196 | helix | exposed |
| MBP | 7 | fkgvdaqgtl | 169 | helix | exposed |

**Table T11.** Secondary structure analysis of proteins with positive hits at 6 or more variable positions in the negative dataset

| **Protein** | **Match** | **Sequence match** | **Position** | **Secondary structure** | **Accessibility** |
|---|---|---|---|---|---|
| DSCAM | 6 | vkaaaasasm | 1196 | helix | exposed |
| ADAM12 | 6 | lkpdavcahg | 459 | loop | exposed |
| FANCD2 | 6 | vkllkisgii | 50 | helix | exposed |
|  | 6 | ikfilhsvta | 283 | helix | exposed |
| Androgen Receptor | 6 | gkvkpiyfht | 911 | loop | exposed |
| KRT8 | 6 | lkgqraslea | 325 | loop | exposed |
| ` | 6 | skfrhaqgtv | 472 | loop | exposed |

**NIRF**

```
                            60        70        80
                          ....|....|....|....|....|....|
Homo sapiens              --CEPVTFEDVTLGFTPEEWGLLDLKQKSL
Pan troglodytes           --CEPVTFEDVTLGFTPEEWGLLDLKQKSL
Canis familiaris          EKEEPVTFEDVILGFTSEEWGLLDLQQKSL
Bos taurus                ---EPVTFEDVALGFTPDEWGKLDLEQKSL
Mus musculus              --HESVKFEDVSLTFTEEEWAQLDFQQKCL
Mus musculus              --HESVKFEDVSLRFTEEEWALLDRQQKCL
Rattus norvegicus         --HESVKFEDVSLTFTKEEWAQLDLQQKCL
```

**TrkA**

```
                                  560       570       580
                          ....|....|....|....|....|....|
Homo sapiens              GKGSGLQGHIIENPQYFS------DACVHH
Pan troglodytes           GKGSGLQGHIIENPQYFS------DACVHH
Canis familiaris          GKGSGLQGHIIENPQYFS------DACVHH
Bos taurus                GKGSGLQGHIIENPQYFS------DACVHH
Mus musculus              GKGSGLQGHIMENPQYFS------DTCVHH
Rattus norvegicus         GKGSGLQGHIMENPQYFS------DTCVHH
Gallus gallus             SKLDGLKSNFIENPQYFC------NACVHH
Danio rerio               GTLDSGLSSFVENPQYFCGIIKDKDMCVQH
```

**TrkB and NTRK2**

```
                                  860
                          ....|....|....|.
Homo sapiens              LQNLAKASPVYLDILG
Pan troglodytes           LQNLAKASPVYLDILG
Canis familiaris          LQNLAKASPVYLDILG
Bos taurus                ----------------
Mus musculus              LQNLAKASPVYLDILG
Rattus norvegicus         LQNLAKASPVYLDILG
Gallus gallus             LQNLAKASPVYLDILG
Danio rerio               LQSLAKASPVYLDILG
```

**TrkC (site 1) and NTRK3 (site 1)**

```
                           610       620       630
                   ....|....|....|....|....|....|...
Homo sapiens       HIVKFYGVCGDGDPLIMVFEYMKHGDLNKFLRA
Pan troglodytes    HIVKFYGVCGDGDPLIMVFEYMKHGDLNKFLRA
Canis familiaris   HIVKFYGVCGDGDPLIMVFEYMKHGDLNKFLRA
Bos taurus         HIVKFYGVCGDGDPLIMVFEYMKHGDLNKFLRA
Mus musculus       HIVKFYGVCGDGDPLIMVFEYMKHGDLNKFLRA
Rattus norvegicus  HIVKFYGVCGDGDPLIMVFEYMKHGDLNKFLRA
Gallus gallus      HIVKFYGVCGDGDPLIMVFEYMKHGDLNKFLRA
```

**TrkC (site 2) and NTRK3 (site 2)**

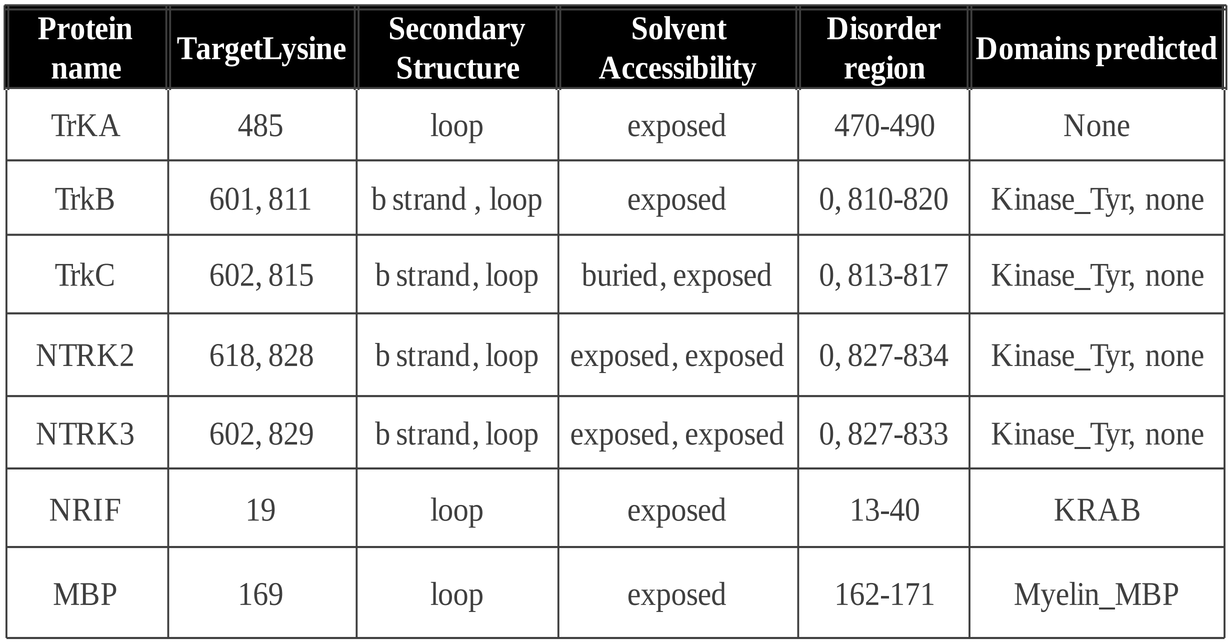
```
                   830       840       850
                   |....|....|....|....|....
Homo sapiens       RLNIKEIYKILHALGKATPIYLDI
Pan troglodytes    RLNIKEIYKILHALGKATPIYLDI
Canis familiaris   RLNIKEIYKVLHALGKAAPIYLDI
Bos taurus         RLNIKEIYKILHALGKATPIYLDI
Mus musculus       RLNIKEIYKILHALGKATPIYLDI
Rattus norvegicus  RLNIKEIYKILHALGKATPIYLDI
Gallus gallus      RLNIKEIYKILHALGKATPIYLDI
```

**MBP**

```
                          310       320
                  ....|....|....|....|....|
Homo sapiens      GFKG--VDAQGTLSKIFKLGGRDSR
Pan troglodytes   GFKG--VDAQGTLSKIFKLGGRDSR
Canis familiaris  GLKG--TDAQGTLSKIFKLGGRDSR
Bos taurus        GLKG--HDAQGTLSKIFKLGGRDSR
Mus musculus      --------------------GRDSR
Gallus gallus     GHKGSYHEGQGTLSKIFKLGGSGSR
Danio rerio       -------SESDELQTIHEHGGAGSE
```

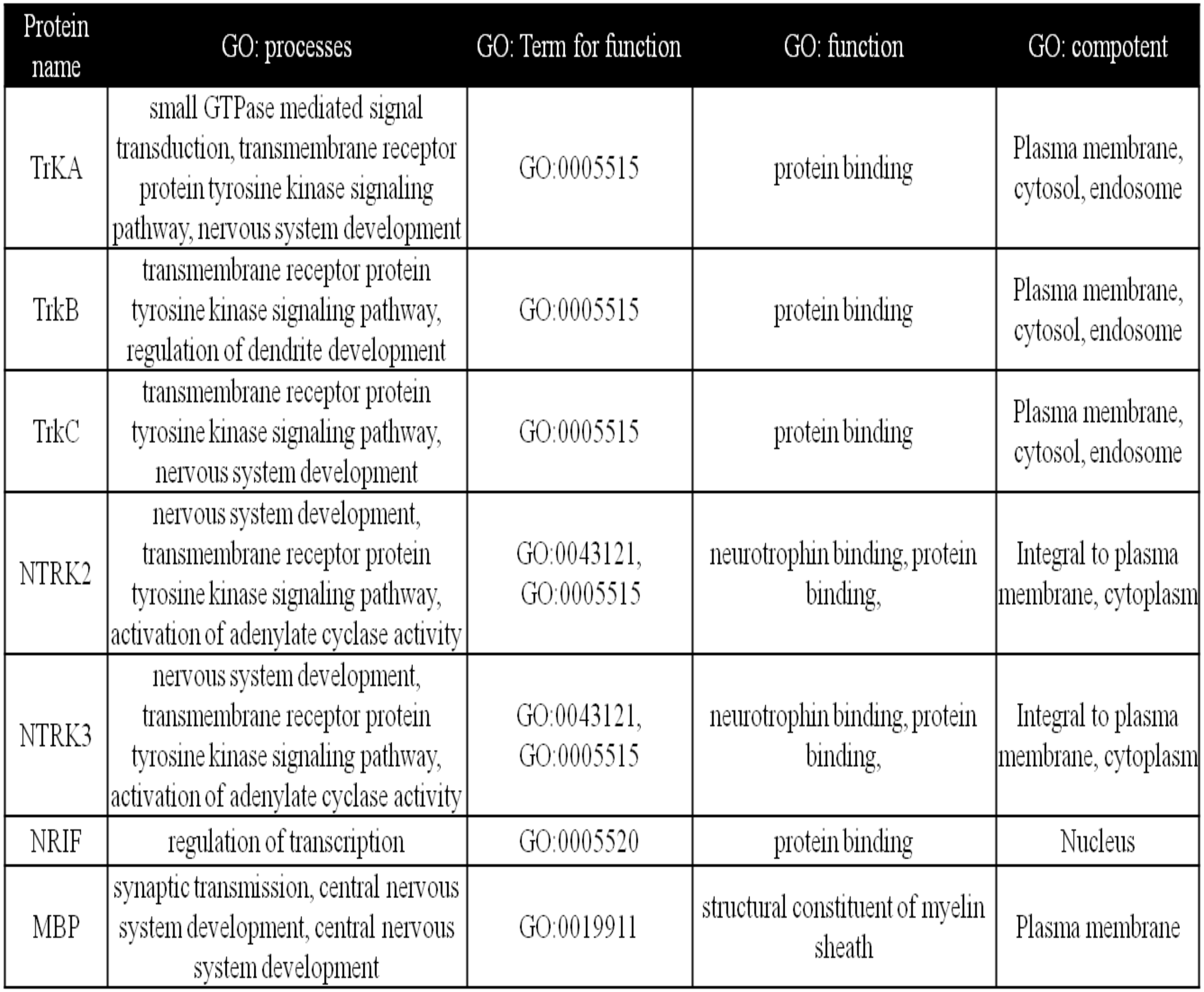
**Figure S2.** Sequence conservation across species at the predicated ubiquitination sites. Proteins NRIF, TRKA, TRKB,TRKC, NTRK2, NTRK3 and MBP had perfect match to the hypothesized motif for TRAF6/p62 ubiquitination.

**Table T12.** Secondary structure analysis of the predicated ubiquitination sites in the high probability proteins with perfect match to the hypothesize motif for TRAF6/p62 ubiquitination



| Protein name | TargetLysine | Secondary Structure | Solvent Accessibility | Disorder region | Domains predicted |
|---|---|---|---|---|---|
| TrKA | 485 | loop | exposed | 470-490 | None |
| TrkB | 601, 811 | b strand , loop | exposed | 0, 810-820 | Kinase_Tyr, none |
| TrkC | 602, 815 | b strand, loop | buried, exposed | 0, 813-817 | Kinase_Tyr, none |
| NTRK2 | 618, 828 | b strand, loop | exposed, exposed | 0, 827-834 | Kinase_Tyr, none |
| NTRK3 | 602, 829 | b strand, loop | exposed, exposed | 0, 827-833 | Kinase_Tyr, none |
| NRIF | 19 | loop | exposed | 13-40 | KRAB |
| MBP | 169 | loop | exposed | 162-171 | Myelin_MBP |

**Table T13.** GO ontology analysis of the predicated ubiquitination sites in the high probability proteins with perfect match to the hypothesize motif for TRAF6/p62 ubiquitination



| Protein name | GO: processes | GO: Term for function | GO: function | GO: compotent |
|---|---|---|---|---|
| TrKA | small GTPase mediated signal transduction, transmembrane receptor protein tyrosine kinase signaling pathway, nervous system development | GO:0005515 | protein binding | Plasma membrane, cytosol, endosome |
| TrkB | transmembrane receptor protein tyrosine kinase signaling pathway, regulation of dendrite development | GO:0005515 | protein binding | Plasma membrane, cytosol, endosome |
| TrkC | transmembrane receptor protein tyrosine kinase signaling pathway, nervous system development | GO:0005515 | protein binding | Plasma membrane, cytosol, endosome |
| NTRK2 | nervous system development, transmembrane receptor protein tyrosine kinase signaling pathway, activation of adenylate cyclase activity | GO:0043121, GO:0005515 | neurotrophin binding, protein binding, | Integral to plasma membrane, cytoplasm |
| NTRK3 | nervous system development, transmembrane receptor protein tyrosine kinase signaling pathway, activation of adenylate cyclase activity | GO:0043121, GO:0005515 | neurotrophin binding, protein binding, | Integral to plasma membrane, cytoplasm |
| NRIF | regulation of transcription | GO:0005520 | protein binding | Nucleus |
| MBP | synaptic transmission, central nervous system development, central nervous system development | GO:0019911 | structural constituent of myelin sheath | Plasma membrane |

**Figure S3.** MATLAB code for MotifMaker program.

```
% Polar 1
O(1) = 'Q';
O(2) = 'Y';
O(3) = 'C';
O(4) = 'S';

% Polar 2
L(1) = 'H';
L(2) = 'D';
L(3) = 'T';

% Hydrophobic
P(1) = 'A';
P(2) = 'L';
P(3) = 'V';
P(4) = 'M';
P(5) = 'G';
P(6) = 'F';
P(7) = 'I';

% Open output file and clean up
```

```
  fid = fopen('Motifout.out','w');
   fclose(fid);


 fid = fopen('Motifout.out','a');
for i1=1:7
       s1 = P(i1);
  for i2=1:7
        s2 = P(i2);
    for i3=1:7
          s3 = P(i3);
     for i4=1:4
            s4 = O(i4);
      for i5=1:7
             s5 = P(i5);
        for i6=1:3
             s6 = L(i6);
          for i7=1:7
             s7 = P(i7);
           motif = [s1,s2,s3,s4,s5,s6,s7];
           % fprintf(fid,' %s, %s, %s, %s, %s, %s,% s\n ', s1,s2,s3,s4,s5,s6,s7);
             fprintf(fid,' %s\n ', motif);


       end
```

```
              end

             end

            end

           end

         end

       end

          %Clear tempSpace from memory

           clear space



          % Lastly CLEAR ALL variables from memory

       fclose('all');

  clear
```

**Figure S3.** MATLAB code for MotifFinder program .

```
% Open output file and clean up
  fida = fopen('MotifsFoundset5.out','w');
   fclose(fida);
  M(201684) = 0;

   fidb = fopen('MotifsFoundset5.out','a');
   fid = fopen('Motifout.out','r');

% Change DataSet.txt to name of current data set
   fidl = fopen('dataset5.txt','r');

 % Input protein sequences for searching
 % Change i1 to equal the number of sequences in the file
 % This is the outside of the big loop

  for i1=1:1

  % This gets the AA sequence, one line at a time
  % Sequence must have no returns
  % Size(A) determines the length of the inputted sequence
  % findstr returns all positions of K in the sequence
  % numel(C) returns the total number of Ks found
```

```
A = fgetl(fidl);

B = size(A);

C = findstr('K', A);

D = numel(C);


% output info about sequence

fprintf(fidb,' Sequence= %d, Length= %d, #_of_k= %d\r', i1, B(2), D);


% sets up null conditions for Motif and Pattern place holders

Best = 'XXXXXXX';

flag = 0;


% this is used to force the first occurance of Pattern to 0 0 0 0 0 0

HH = 'AAAAAAA';

Pattern = HH==Best;


for i2=1:D

 % Identify the characters in the appropriate Motif positions

 % Then merge until a single character array (Test)

 % First check to see if the last K is too close to the end to have a full Motif


  EE = B(2)-C(i2);

  Best = 'XXXXXXX';
 flag = 0;
    cc = 0;
% this is used to force the first occurance of Pattern to 0 0 0 0 0 0 0
    HH = 'AAAAAAA';
   Pattern = HH==Best;

  % The 8 is the mininum number of places from the end of the sequence
  % where a K could occur and still be a full Motif
  % if condition is true then search of motifs is allowed
  if EE >= 8

   % This is the actual search
   % The P() parts pull out the chracters from the actual sequence

   P1 = A(C(i2)-1);
   P2 = A(C(i2)+1);
   P3 = A(C(i2)+4);
   P4 = A(C(i2)+5);
   P5 = A(C(i2)+6);
   P6 = A(C(i2)+7);
   P7 = A(C(i2)+8);

   % This merges the characters into a motif for testing
   R = [P1,P2,P3,P4,P5,P6,P7];
   flag = 0;
   for i3=1:201683
   % for i3=1:10
   % Now pull a test motif and test
   M = fgetl(fid);
   x=M;
   y=R;
   Test = M(3:9)==R;

   correct = sum(Test);
   if correct > flag

       flag=flag+1;
       Best = M;
       YY = R;
       Pattern = Test;
       cc=correct;
  end
 end
    frewind(fid);
   fprintf(fidb,'%d, %d, %d, %d, %d, %d, %d, %d, %s, %d, %d, %s\r', i2, Pattern(1),
 Pattern(2), Pattern(3), Pattern(4), Pattern(5), Pattern(6), Pattern(7), Best, cc, C(i2), YY);
```

```
      end
   end
 end


  finish = 'Finished'
 %Clear tempSpace from memory
    clear space


   % Lastly CLEAR ALL variables from memory
fclose('all');
clear
```