International Journal of Pharmaceutical and Life Sciences
ISSN 2305-0330
Volume 2, Issue 1: January 2013

INVESTIGATION OF A PROTEIN ENCODED BY A GENE OF *MUS MUSCULUS* AND RELATING IT TO AN ORTHOLOGOUS PROTEIN IN *HOMO SAPIENS*

Avijit Mallick

Department of Biochemistry and Molecular Biology, University of Melbourne, Australia

Abstract: The gene of Mus musculus was cloned into plasmid vector (pYEUra3) of E.coli from the cDNA library which was then verified using colony PCR. The plasmid DNA was then purified and digested with EcoRI to check the presence of an insert. PICOGREEN assay, HindIII digestion and absorbance were used to quantify the DNA. The purified DNA was then sequenced and searched using different databases to find a match. Finally, the protein of interest was identified to be Sentrin-specific protease 2(SUMO-1 protease 1) which processes SUMO precursor into their mature form and also causes deSUMOylation from their target substrate.

Keywords: SENP2, SUMOylation, Protein

Introduction: The three techniques that have revolutionized research in biochemistry are cDNA libraries, advanced sequencing technology and bioinformatics. First of all, cDNA libraries have made life a lot easier for scientists. It has been very useful for characterizing the structure and function of newly synthesized genes.[1] cDNA is created from the mRNA using the enzyme reverse transcriptase which is then converted into double stranded cDNA using RNase-H and DNA polymerase. The dscDNA is then cloned into a vector which then replicates and makes multiple copies of the cDNA. cDNA library is made from all the mRNA expressed so therefore it contains all the genes of an organism. In such a way, the cDNA library for *Mus musculus* was produced from which the gene was cloned into *Ecoli* bacteria.

Secondly, DNA sequencing technology has revolutionized many different part of science like archaeology, anthropology, genetics, biotechnology, molecular biology, forensic science and others. There are four well DNA sequencing technologies among which Sanger method and its most important variants (enzymatic methods) have been used in our experiment. This method which changed the field of genetics was well known as chain termination method or the dideoxynucleotide method. Here, DNA polymerase carries out catalytic polymerization of dNTP until a ddNTP is incorporated where the polymerization stops. The idea is to produce strands of different lengths that differ by one nucleotide which can be detected by fluorescence attached to ddNTP and used to sequence the DNA [2].

And thirdly, biological questions are now answered using computational approaches represented by a brand new emerging area of science called bioinformatics. This approach is changing the way how basic science is done to a more efficient guided design. Bioinformatics is now playing a very significant role to solve biological problems by providing an explosion of sequence and structures to researchers.

Correspondence to Author:
Email: avijitsunshine@gmail.com

Most importantly, this approach will face the challenge with the aid in gene discovery and molecular designing, site directed mutation and researches that will reveal unknown relationship between the structure, function of genes and proteins [3]. Bioinformatics has been used in our experiment to find the best match of coded proteins for our gene of interest from *Mus musculus.*

The overall flowchart of our experiment used all the three described techniques above. The DNA from the cDNA library of *Mus musculus* was cloned into *E.coli* which was then extracted and sequenced using the Big-dye terminator method. Our sequenced DNA was then searched in the three Bioinformatics databases to find the best the match of protein encoded by this sequence. Finally, the protein of best match to be found was SUMO-1 protease-1 which is also found in *Homo sapiens* with same structure and function. SUMO specific proteases processes SUMO precursor to reveal C-terminal glycine residue that is linked to lysine side chains in target proteins. This protein also cleaves SUMO from modified protein. Eight genes have been identified for this protein [4] of which SENP1, SENP2 and SENP3 are SUMO specific in human[5]. In our study we focus on SENP2 from *Mus musculus* and try to understand its function in *Homo sapiens.*

Materials and Methods:

COLONY PCR identification of insert: The colony of *E.coli* containing plasmid (pYEUra3) was inoculated with cDNA insert and allowed to grow overnight. One of the colony (31) was then selected and prepared for colony PCR reaction using the oligonucleotide primers (**YEURAfor**: CACACTGTGGTAGAGC; **YEURArev**: CTCACAAATTAGAGC), polymerase and other reaction components as described in [6].

Plasmid DNA preparation: The plasmid DNA was then purified from the clones using Wizard Plus Miniprep DNA purification system as directed in the protocol given by (Promega Corporation U.S.A)

PICOGREEN assay: The purified DNA was quantified by fluorescence. In this assay standard DNA (2µg/mL) was used. Standard DNA volume of 0, 44, 88, 175 and 263µL were taken in separate tubes to which 315, 217, 227, 140 and 52µL TE buffer was added. Our sample was prepared with 17.5µL of DNA and 298µL of TE buffer.315µL Picogreen Dye was then added to all the tubes and assayed on the Fluostar Omega plate following the protocol of the PICOGREEN program.

Restriction digestion: *HindIII* digestion was carried out with 2µL NEB buffer 2, 12µL water, 5µL DNA and 1µL *HindIII. Eco*RI digestion was carried in slightly different conditions with 2µL NEB buffer, 2µL *Eco*RI, 6µL water and 10µL DNA.

Agarose gel electrophoresis: Gel electrophoresis of colony PCR, *HindIII* digestion and *Eco*RI digestion of DNA was carried out according to the methods outlined in [7]. Colony PCR and *HindIII* digestion electrophoresis was carried out in 1% agarose gel whereas that of *Eco*RI was carried out on 1.2% agarose gel.

Sequencing PCR: The plasmid DNA was then sequenced by cycle sequencing using the Big Dye Terminator Reaction protocol.**(Automated DNA cycle sequencing)** from [8] using the primer (**YEURAfor**: CACACTGTGGTAGAGC).

Bioinformatics: The whole length DNA(vector + insert) sequence was edited to remove the vector sequence using the *Eco*RI site(GAATTC). A Blast search was then carried out

of the sequence of interest using ExPASy and a match was found out from all six translated proteins. In addition, another Blast of the sequence was carried out using NCBI Blatx and NCBI ORFfinder was used. The proteins from all the three databases were then matched to validate the result.

Results:

A. Gel Electrophoresis of colony PCR:

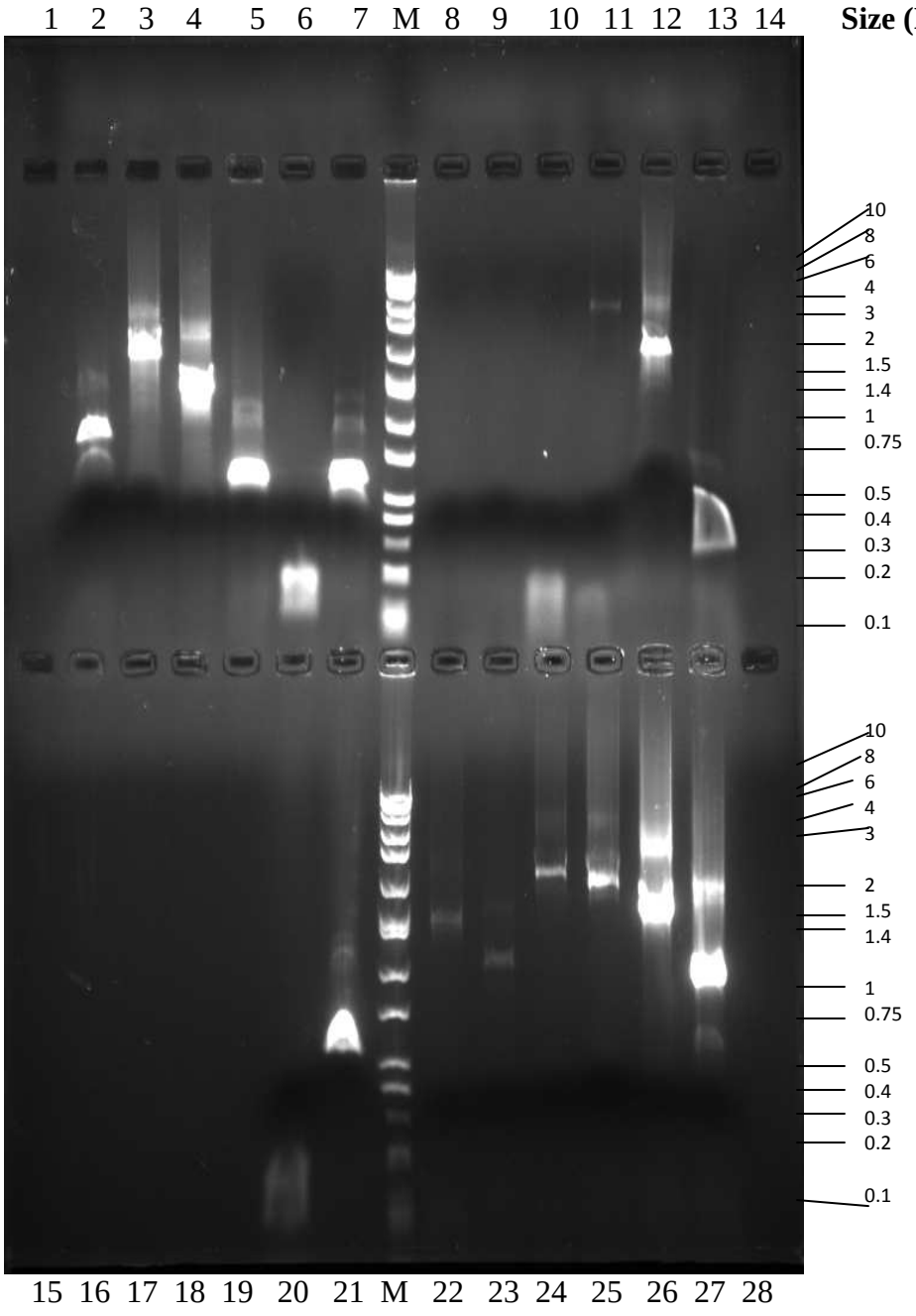


Figure 1:- Gel Electrophoresis of colony PCR

This was carried out to check the presence of an insert in the colonies and to determine the success of colony PCR. Colony 31 in **lane 21** showed one band of **631bp** which confirmed the presence of an insert in the DNA. The molecular weight for the insert was expected in the range of 600bp to 4000bp and that without any insert of around 100bp. 77% of the colony had an insert between 500 to 3000bp. This result was very close to our expectation and satisfactory as 70% of the insert were expected to have an insert. The plasmid DNA was then purified and digested with *EcoRI* to confirm the presence of an insert which is presented below.

B. Gel electrophoresis of plasmid DNA after digestion with *E*coRI:

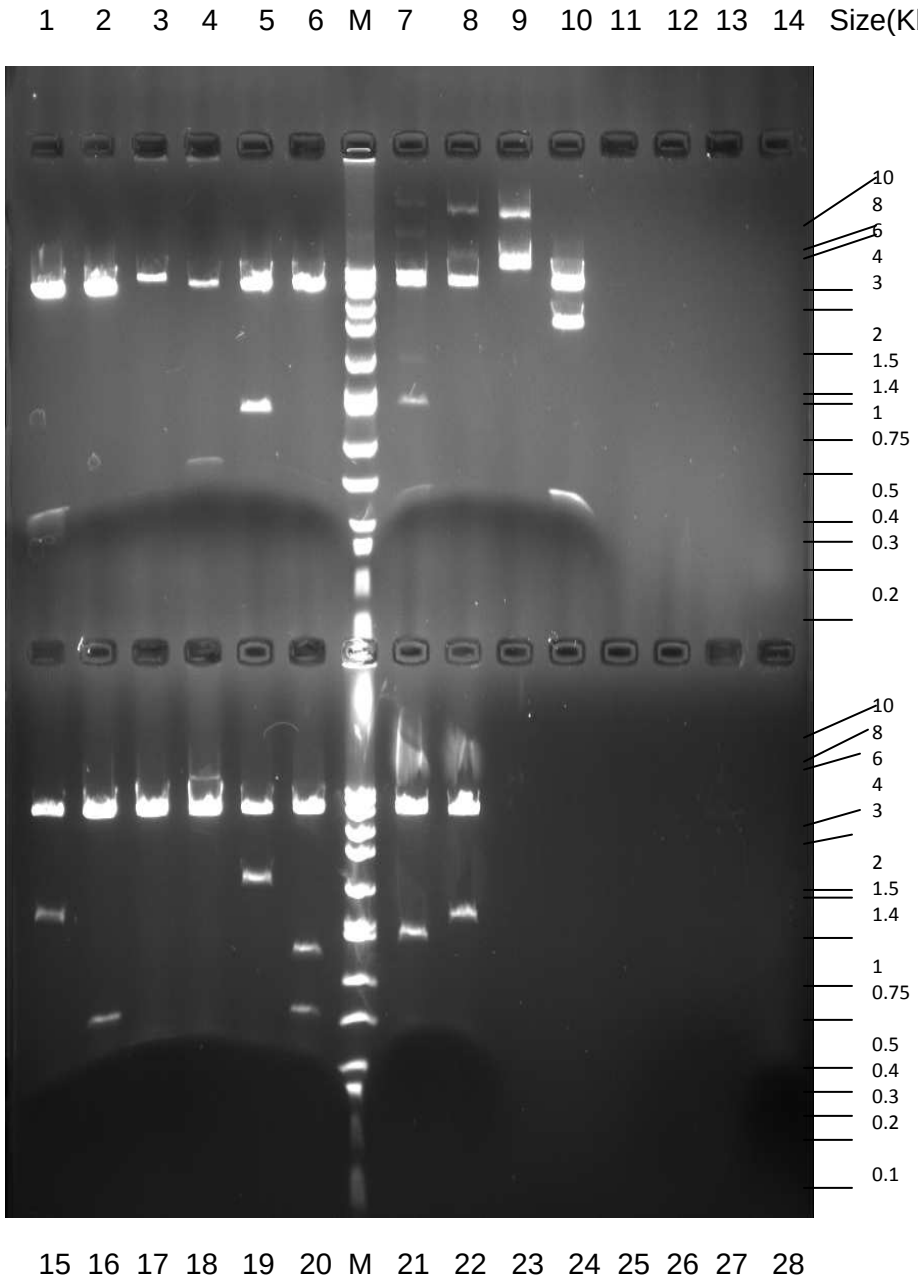


Figure 2:- Gel electrophoresis of plasmid DNA after digestion with *E*coRI

There was only one band seen in lane 17(colony 31) unexpectedly but, had a possibility of the smaller band being surmounted by the Bromophenyl blue which also migrated around 300bp to 500bp. So, my result was not consistent with the colony PCR gel electrophoresis. In contrast, as a whole class 61% (expected 77% from colony PCR) of them had an insert ranging between 3000bp to 500bp which was less than expected and not consistent with the colony PCR electrophoresis result. This would have been due to the unsuccessful digestion of the DNA with *Eco*RI. In the next part, the purified DNA was quantified to know the exact concentration which was then used to calculate the volume of plasmid DNA required to give 300ng of DNA for the sequencing reaction.

C. Concentration of DNA from colony 31 determined using three methods:

NO	METHOD	Concentration(ng/µL)
1	Absorbance (a)	34.25
2	*HindIII* digestion (b)	20.00
3	PICOGREEN assay	12.90

(a) Calculated using Beer Lambert's law and ε =200%$^{-1}$cm^{-1}
(b) Calculated by comparing the intensity with the marker of known concentration.

Table 1: Concentration of DNA from colony 31 determined using three methods

Among the three methods concentration determined by PICOGREEN dye was the most reliable. Concentration determined by *HindIII* digestion was not reliable as it was estimated on the band intensity of the marker which was really hard to distinguish. Data obtained by absorbance showed furthest deviation from the other values and was the least reliable as contaminating nucleic acids also gave an absorbance at 260nm. Based on the PICOGREEN result 23µL (300ng) of DNA was transferred for the sequencing PCR reaction .The sequenced DNA was then used to search the best possible match of protein in three different databases.

D. Top five searches from the blast search using Expasy:

AC	DESCRIPTION	S	E
Q91ZX6	SENP2_MOUSE Sentrin-specific protease 2 (EC 3.4.22.68)	164	8e-39
Q91ZX6-2	Isoform 2 of Sentrin-specific protease 2 OS=Mus musc.	164	8e-39

AC	DESCRIPTION	S	E
Q91ZX6-3	Isoform 3 of Sentrin-specific protease 2 OS=Mus musc...	164	8e-39
Q9EQE1	SENP2_RAT Sentrin-specific protease 2 (EC 3.4.22.68)	157	7e-37
D3ZF69	RAT Sentrin-specific protease 2 [Senp2] [Rattus norveg...	157	7e-37

(AC)- Accession number (Denoted by 6 alphanumerical characters) is a stable identifier of UniProt KB entries.
S(Score)- It denotes the score of an alignment calculated as the sum of substitution and gap score.
E- Expectation value represents number of different alignments with scores equivalent to or better than S that is expected to occur.

Table 2: The top five searches from the blast search using Expasy

Here the protein database was searched using the translated nucleotide sequence. The protein with the best match had the highest S value and lowest E value. Although the S value was not that high the E value was low as expected. 8e-39 was quite a low value which showed that the protein match by chance was very low. E value is the most important value to look at compared to S. The top three searches were the same protein but the isoform of each other. Finally we knew our protein which was Sentrin specific protease 2 also known as SUMO-1 protease-1. Then we found out the alignment match between the query and subject sequence which is presented below.

Alignments

sp	Q91ZX6 SENP2_MOUSE	**Sentrin-specific protease 2 (EC 3.4.22.68) (Axam2) (SUMO-1 protease 1) (SuPr-1) (SUMO-1/Smt3-specific isopeptidase 2) (Smt3ip2) (Sentrin/SUMO-specific protease SENP2) [Senp2] [Mus musculus (Mouse)]**	588 AA

Score = 164 bits (414), Expect = 8e-39

Identities = 77/77 (100%), Positives = 77/77 (100%)

Query: 6 GYNRRPSGRRHSKSNPESSLTWKPQEQGVTEMISEEGGKGVRRPHCTVEEGVQKD EREKY 65

GYNRRPSGRRHSKSNPESSLTWKPQEQGVTEMISEEGGKGVRRPHCTVEEGVQKD EREKY

Sbjct: 158 GYNRRPSGRRHSKSNPESSLTWKPQEQGVTEMISEEGGKGVRRPHCTVEEGVQKD EREKY 217

Query: 66 RKLLERLKEGAHGSTFP 82

RKLLERLKEGAHGSTFP

Sbjct: 218 RKLLERLKEGAHGSTFP 234

Table 3:- The alignment sequence is showed with identities and positives values

This shows the matched sequence of our query with the sequence of the protein (Sentrin specific protease 2).Our sequenced DNA matched with an identity of 100% and positives of 100% with the subject sequence. Although the protein consisted of 588AA the match was between 158 to 234 AA. For further confirmation whether our DNA really coded for this protein two more databases were searched.

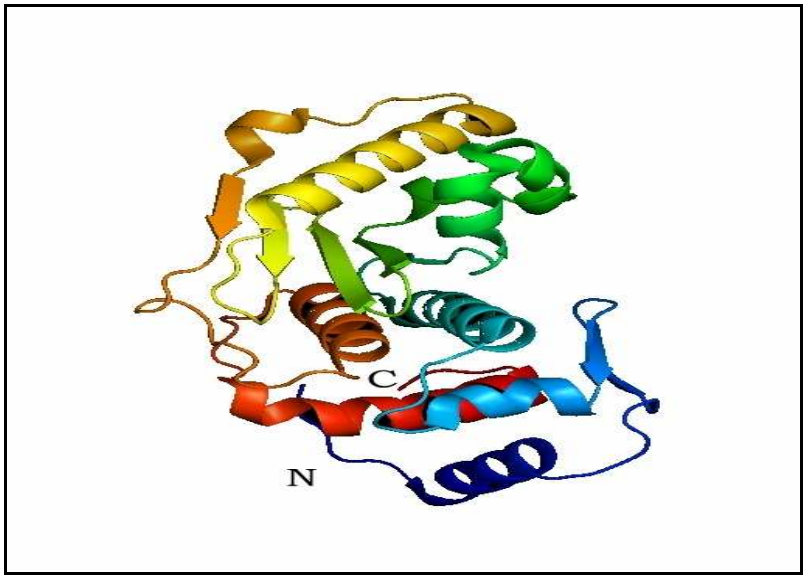
E. Results from three databases with the score and E values:

DATABASE		Description	SCORE	E
Expasy	Q91ZX6	Sentrin-specific protease2	164	8e-39
NCBI(blastx)	AA0279021	SUMO1 protease-1	164	9e-45
NCBI(ORFfinder)	AA0279021	SUMO1 protease-1	97.1	2e-22

Table 4:- Results from three databases with the score and E values

The top match from each search is shown in each case. The protein is the same in all the three databases but there E values were quite different. Expasy and NCBI (blastx) have the same S value except the NCBI(ORFfinder) All the searches gave a reasonably small E value and high S value. As in each case, the top match was the same we confirmed our protein of interest to be Senp2. The ribbon structure of the molecule was then prepared using Pymol software.

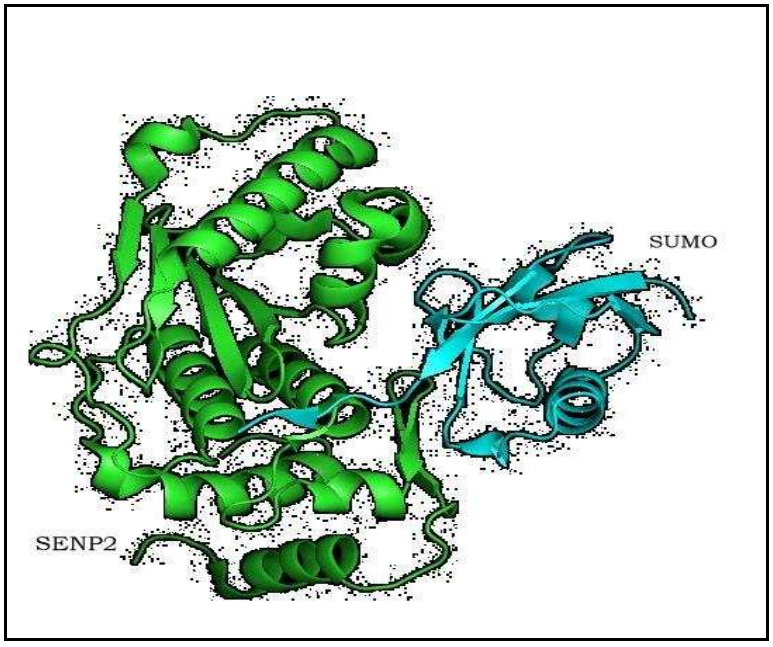
F. Ribbon structure for Senp2 catalytic domain:



PDB code:-1TH0 [9]
Graphics prepared using Pymol

FIGURE 3: Showing the ribbon structure for Senp2 catalytic domain

The chain from N-terminal starts with violet and follows the rainbow and ends in red with the C-terminal.



PDB code:-1TGZ [9]
Graphics prepared by Pymol

FIGURE 4: Showing the ribbon structure of SUMO-1 Senp2 complex

Discussion:

When the sequencing of the DNA was complete it was modified to remove plasmid sequence from both the ends. The sequence was then translated into amino acid sequence and searched for the perfect combination of codons coding the protein of interest. The amino acid sequence was then blast searched to find the best match with low E (Expected) value and high S (Score) value. Three databases were used to confirm the identification of the protein (as shown in Table 4).Expasy blast showed the protein to have closest match with Senp2 with Score of 164 and E of 8e-39. The values were as expected with high S and low E value. The second database search which was NCBI (blast x) gave the same score value of 164 but a better E value of 9e-45. And finally, the third database search with NCBI (ORFfinder) gave different S and E values of 97.1 and 2e-22.But, all the three databases showed the best possible match as **SUMO-1 protease-1(Senp2)** which confirmed the protein of our interest (as shown in Table 4).

SUMO (Small ubiquitin like modifier) when linked to proteins by covalent bond alters the fate of the target proteins even though it might get deconjugated in a very short period time. This ubiquitin like modifier changes the properties of the target proteins and makes the proteome very complex in eukaryotic cells [5]. Sumo was first identified in mammals where it was bounded to GTPase activating protein RANGAP1. Furthermore, it was found to have very low identity of 20% with ubiquitin [10, 11]. SUMOylation pathway affects many biological processes like nuclear metabolism and cell cycle progression [9] and is important for cell viability in yeasts, nematodes and higher eukaryotes [12, 13, 14].

Vertebrates have three paralogues of the SUMO family which are SUMO-1, SUMO-2 and SUMO-3 [5, 9].

Senp2 (SUMO-1 protease-1) is our protein of interest shown in (Figure 3) which is an important enzyme in the deSUMOylation pathway. It detaches SUMO from the target protein in an ATP dependant manner.

Senp2 function, structure and its relation to any diseases:

Function: This protein is a protease that that plays a major role in SUMO pathway: Senp2 processes SUMO-1, SUMO-2 and SUMO-3 into their mature form. SUMO is proteolytically processed to give a conserved C-terminal Gly-Gly motif. It also deconjugates SUMO-1, SUMO-2 and SUMO-3 from their targeted protein [15]. The attachment of SUMO (Small ubiquitin-like modifiers) to proteins is a reversible posttranslational modification (as shown in Figure 5) that controls protein's function, sub cellular localization and expression. The SUMO proteases are very important as they deconjugate the modified proteins and maintain the level of SUMOylated and un-SUMOylated proteins for normal physiology [16]. In addition, Senp2 have three isoforms which are found in nuclear pore complex, cytoplasmic vesicle and nucleus [15].

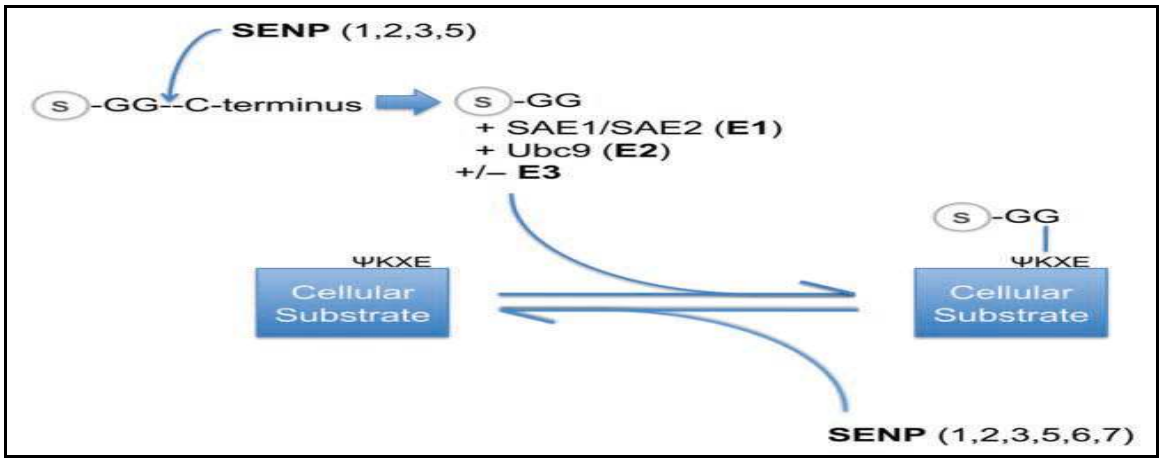


Figure 5: The process of SUMOylation and deSUMOylation showing the involvement of Senp2 (Bawa-Khalfe and Yeh, 2010)

Structure: Senp2 consists of 588AA and have an Ulp-like catalytic domain at the C-terminal. It has two sub domains with α-helix rich segment at N-terminal and five stranded β sheet surrounded by two α-helices forming the core of the C-terminal sub domain. Senp2 active site consists of a catalytic triad of amino acid residues that includes Cys548; His478 and Asp495.Structural reorganization takes place during the binding of Senp2 with SUMO-1. His478 rotates by 180 degree in the Senp2-SUMO-1 complex which causes the imidazole ring to point directly towards Asp495 and away from Cys548. Senp2 is very specific and can distinguish SUMO-1 from SUMO-2 and SUMO-3 with structural rearrangement taking place in response to SUMO-1 binding. SUMO-1 crystal structure was first determined through Senp2-SUMO-1 complex as shown in Figure 4 [9].

Disease caused by protein: Experiments on other isoforms have showed that SUMO protease can down-regulate CTNNB1 levels and there by modulate the Wnt-pathway. This suggests that Senp2 may affect the Wnt signaling pathway through β-catenin degradation. Wnt signaling is a very important pathway where β-catenin is a transcription factor regulating the genes related to cell proliferation.

SENPs are also found to have altered expression in carcinomas. It has been confirmed in previous studies that SENP levels disrupts the homeostasis of SUMO and causes cancer development and progression [16].

Senp2 is highly expressed in testis of mouse but is also detected in brain, heart and thymus. The same protein is also found in *Homo sapiens* (Mammals) [9] which have the same function. It is not possible to check the function of this protein with mutation in humans so mouse can be used as a very effective model for human. How do the mutations at the binding site of Senp2 affect the whole process of desumoyltion and vice versa? How might it be related to any major diseases like cancer in humans? Experiments should be directed to check the affect of several mutations on Senp2 that affects the process of deSUMOylation and led to diseases like cancer.

Acknowledgement:

Leon Helfenbaum
Department of Biochemistry and molecular biology
University of Melbourne

References:

1) Zhu et al. (2001) Reverse transcriptase template switching: A SMART Approach for Full-Length cDNA library construction*, BioTechniques* 30 892-897.

2) Franca, Camilho and Kist (2002) A review of DNA sequencing Technologies, Quarterly Reviews of Biophysics 35 169-200.

3) Baxevanis. (2001) Bioinformatics and the internet, Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Second Edition (Chapter: 1), 1-17

4) Yeh, E.T., Gong, L., and Kamitani, T. (2000). Ubiquitin-like proteins: new wines in new bottles. Gene *248*, 1–14.

5) Hay.(2005) SUMO: A History of Modification, *Molecular Cell* 12

6) Güssow, D. and Clackson, T. (1989) Direct clone characterisation from plaques and colonies by the polymerase chain reaction, Nucleic Acids Res. 17, 4000 1-12

7) Sharp, P.A., Sugden, W. and Sambrook, J. (1973) Detection of two restriction endonuclease activities in Haemophilus parainfluenzae using analytical agarose-ethidium bromide electrophoresis, Biochemistry 12, 3055-3063

8) Rosenblum, B.B. *et al*. (1997) New dye-labelled terminators for improved DNA sequencing patterns, Nucleic Acids Res. 25, 4500 4504

9) Reverter and Lima. (2004) A Basis for SUMO Protease Specificity Provided by Analysis of Human Senp2 and a Senp2-SUMO Complex, *Structure* 12 1519-1531.

10) Mahajan, R., Delphin, C., Guan, T., Gerace, L., and Melchior, F.(1997). A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2. Cell *88*, 97-107.

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

12) Fraser, A.G., Kamath, R.S., Zipperlen, P., Martinez-Campos, M.,Sohrmann, M., and Ahringer, J. (2000). Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. Nature *408*, 325–330.

13) Hayasi, T., Seki, M., Maeda, D., Wang, W., Kawabe, Y., Seki, T.,. Saitoh, H., Fukagawa, T., Yagi, H., and Enomoto, T. (2002). Ubc9 is essential for viability of higher eukaryotic cells. Exp. Cell Res. *280*, 212–221.

14) Seufert, W., Futcher, B., and Jentsch, S. (1995). Role of a ubiquitin- conjugating enzyme in degradation of S- and M-phase cyclins. Nature *373*, 78–81.

15) Nishida, T., Tanaka, H., and Yasuda, H. (2000). A novel mammalian Smt3-specific isopeptidase 1 (SMT3IP1) localized in the nucleolus at interphase. Eur. J. Biochem. 267, 6423–6427

16) Bawa-Khalfe and Yeh (2010) SUMO Losing Balance: SUMO Proteases Disrupt SUMO Homeostasis to Facilitate Cancer Development and Progression, *Genes & Cancer* 1(7) 748-752.

17) National centre for Biotechnology Information (2009) (http://www.ncbi.nlm.nih.gov) Accessed 20 August 2012.

18) SIB Swiss Institute of Bioinformatics (2011) Bioinformatics Research Portal (http://www.expasy.org) Accessed 20 August 2012.