Study of Small Ubiquitinin like Modifiers and Chaperonic Signaling in Non-Insulin Dependent Diabetes Mellitus

A Major Project dissertation submitted in partial fulfilment of the requirement for the degree of

Master of Technology
In
Bioinformatics

Submitted by

Ravi Kumar Tomar

(2K11/BIO/15)

Delhi Technological University, Delhi, India

Under the supervision of

Dr.Pravir Kumar

Associate Professor

Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
ShahbadDaulatpur, Main Bawana Road,
Delhi-110042, INDIA

DELHI TECHNOLOGICAL UNIVERSITY, New Delhi

CERTIFICATE

This is to certify that the M. Tech. dissertation entitled **“Study of small ubiquitin like modifiers and chaperonic signalling in non-insulin dependent diabetes mellitus”**, submitted by **Ravi Kumar Tomar (2K11/BIO/15)** in partial fulfilment of the requirement for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate’s own work carried out by him under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date:

Dr. Pravir Kumar
Associate Professor
Department of Bio-Technology
Delhi Technological University
(Formerly Delhi College of Engineering, University of Delhi)
Adjunct Faculty Tufts University School of Medicine, Boston, MA, USA

ACKNOWLEDGEMENT

I express my sincere thanks and deepest gratitude to my guide ***Dr Pravir Kumar*** *Department of Biotechnology, Delhi Technological University for giving hisunstined, sagacious guidance, advice and supervision throughout my thesis work. I am highly impressed to his Intelligence, well organized and enthusiastic approach towards goal-oriented research. I am privileged to have been his student. I am greatly thankful to her for providing excellent laboratory facilities, perfect scientific environment as well opportunities to explore scientific world.*

I am greatly thankful to all respected faculty of the Department of Biotechnology, DTU for their efficient Teaching, generous support and providing me clear concept in Biotechnology.

*Mere words are not enough to express my feelings towards my Parents***.** *It was their dream and guidance that made me strong enough to pursue further studies.*

RAVI KUMAR TOMAR

M. Tech. Bioinformatics

Batch: 2011-2013

Roll no.: 2K11/BIO/15

DECLARATION

I hereby declare that the work, which is being presented in project entitled **“Study of small ubiquitin like modifiers and chaperonic signalling in non-insulin dependent diabetes mellitus”**in partial fulfilment of the requirement for the award of MASTER OF TECHNOLOGY in BIOINFORMATICS degree, is an authentic record of my own work carried out under the supervision of Dr. Pravir Kumar, Associate Professor, Department of Biotechnology, Delhi Technological University.

The project was undertaken as a part of academic curriculum according to the University rules and norms and it has not commercial interest and motive, it is my original work. It is not submitted to any other organisation for any purpose.

Ravi Kumar Tomar

Roll No. 2K11/BIO/15

TABLE OF CONTENTS

TOPIC **PAGE NO.**

LIST OF FIGURES ….. 8

LIST OF TABLES ….. 8

LIST OF ABBREVIATIONS ….. 9

1. ABSTRACT ….. 10

2. INTRODUCTION ….. 11
    - 2.1. Diabetes Mellitus
        - 2.1.1. Classification ….. 11
    - 2.2. Motivation ….. 11
    - 2.3. Proposed model ….. 12
    - 2.4. Thesis Outline ….. 13

3. REVIEW OF LITERATURE ….. 14
    - 3.1. Non-Insulin Dependent Diabetes Mellitus (NIIDM) ….. 14
        - 3.1.1. Complications ….. 14
        - 3.1.2. Pathogenesis ….. 15
    - 3.2. The Chaperonic System: Physiology and Pathology ….. 16
        - 3.2.1. Chaperoning system ….. 16
        - 3.2.2. Molecules, cells and tissues of the chaperoning system ….. 16
        - 3.2.3. Classifications of molecular chaperones ….. 16
    - 3.3. Heat shock Protein ….. 19
        - 3.3.1. Oligomerization ….. 19
        - 3.3.2. Functional regulation by phosphorylation ….. 20
    - 3.4. Extracellular-signal regulated kinases (ERK) ….. 21
        - 3.4.1. MAPK/ERK pathway ….. 21
        - 3.4.2. Types of ERKs ….. 21
    - 3.5. Small Ubiquitin like Modifier (SUMO) proteins ….. 23
        - 3.5.1. Structure ….. 23
        - 3.5.2. Functions ….. 24
        - 3.5.3. Benefits of SUMO platform ….. 24

4. METHODOLOGY ….. 25
    - 4.1. Protein-protein interaction ….. 25
        - 4.1.1. STRING database ….. 25
        - 4.1.2. VisANT: shortest pathway analysis ….. 26

4.2. Physico-chemical properties: PROTPARAM ….. 27
4.3. Protein Variations: SWISSvar ….. 28
4.4. Multiple sequence alignments: Clustal Omega ….. 29
4.5. SUMO sites: SUMOsp ….. 30
4.6. Domain and motif: PROSITE ….. 30
4.7. Binding sites in protein: 3DLigandSite ….. 31
4.8. Virtual Screening: FINDSITE ….. 31
4.9. Protein-ligand docking ….. 32
4.9.1. PatchDock ….. 32
4.9.2. SwissDock ….. 33

5. RESULTS ….. 34
5.1. Comparative study of SUMO proteins ….. 34
5.2. Variations among SUMO proteins ….. 36
5.2.1. SUMO1 variants ….. 36
5.2.2. SUMO2 variants ….. 36
5.2.3. SUMO3 variants ….. 37
5.2.4. SUMO4 variants ….. 38
5.3. Protein-protein interaction network ….. 39
5.3.1. Shortest pathway Analysis ….. 42
5.4. Multiple Sequence Alignments of proteins ….. 43
5.4.1. Between all four SUMO proteins ….. 43
5.4.2. Between SUMO proteins, GSK3, ERK2, AKT2 and HSP27 ….. 46
5.4.3. Analysis ….. 48
5.5. SUMO binding sites prediction ….. 49
5.5.1. Analysis ….. 51
5.6. Motif and domain analysis ….. 52
5.7. Physico-chemical parameters computation ….. 53
5.7.1. Analysis ….. 53
5.8. Binding sites prediction ….. 54
5.8.1. SUMO1 binding sites ….. 54
5.8.2. SUMO2 binding sites ….. 55
5.8.3. SUMO3 binding sites ….. 56
5.8.4. SUMO4 binding sites ….. 57
5.8.5. Analysis ….. 58
5.9. Virtual Screening of ligand using ZINC database ….. 59
5.10. Protein-ligand docking ….. 62
5.10.1. Analysis of docking results ….. 63

6. CONCLUSION ….. 67

7. DISCUSSION AND FUTURE PERSPECTIVE ….. 68

8. REFERENCES ….. 69

9. APPENDIX ….. 72

9.1.OMIM ….. 72
9.2. ZINC Database ….. 72
9.3. KEGG ….. 73
9.4. DrugBank ….. 73
9.5. Jmol ….. 73
9.6. Lipinski Filters ….. 74
9.7.PyMOL ….. 74

LIST OF FIGURES AND TABLES

List of Figures:

1. Proposed model ... 12
2. Pathogenesis of Type-2 Diabetes Mellitus ... 15
3. Functional regulation of HSP-27 ... 20
4. Proteins involved in the serine/threonine protein kinase Akt pathway and in the ERK1/2 pathway ... 22
5. Benefits of SUMO platform ... 24
6. Protein-protein interaction network of SUMO proteins ... 39
7. ZINC Accession no. and structures of top 10 virtually screened compounds ... 60
8. PyMol images of docked ligands within binding sites of SUMO1 receptor ... 64

List of Tables:

1. Classification of HSP-chaperones according to molecular weight ... 16
2. Places at which chaperones reside and work ... 17
3. Comparative study of SUMO proteins ... 35
4. Score values in multiple sequence alignments ... 44
5. SUMO binding sites prediction on key proteins involves directly or indirectly in glucose metabolism ... 51
6. Predicted Features and Patterns of SUMO proteins ... 52
7. Physico-chemical parameters of proteins ... 53
8. Top 20 virtually screened compounds against SUMO1 receptor ... 59
9. Docking results of top 10 virtually screened ligands with SUMO1 protein. ... 62

LIST OF ABBREVIATIONS

SUMO	Small ubiquitin-like modifier protein
HSP27	Heat shock protein
AKT1/2	RAC-beta serine/threonine-protein kinase
ERK1/2	Mitogen-activated protein kinase 1/2
RANGAP1	Ran GTPase activating protein 1
MDM2	E3 ubiquitin-protein ligase Mdm2
DAXX	Death-domain associated protein
UBA2	Ubiquitin-like modifier activating enzyme 2
TP53	Tumour protein p53
SENP2	SUMO1/sentrin/SMT3 specific peptidase 2
PIAS4	Protein inhibitor of activated STAT, 4
USP25	Ubiquitin specific peptidase 25
PML	Promyelocytic leukaemia
UBE2I	Ubiquitin-conjugating enzyme E2I
G6PC2	Glucose-6-phosphatase, catalytic, 2
FOXC2	Forkhead box protein
DPP-4	Dipeptidyl-peptidase-4
HNF4A	Hepatocyte nuclear factor 4-alpha
IDE	Insulin degrading enzyme
RANBP2	RAN binding protein2

Study of small ubiquitin like modifiers and chaperonic signalling in non-insulin dependent diabetes mellitus

Ravi Kumar Tomar
Delhi Technological University, Delhi, India

ABSTRACT

Small ubiquitin-related modifier (SUMO) family proteins function in post-translational modifications of proteins by covalently attaching to them. This helps in modification of many proteins involve in diverse cellular processes, like nuclear transport, transcriptional regulation, signal transduction and maintenance of genome integrity. An enzyme pathway which is related to the ubiquitin pathway control the attachment between SUMO and proteins.

Heat shock protein 27 (HSP27) helps stressed cells to survive in adverse conditions by accumulating in them. HSP27 function in the ubiquitination process is already known. Here, we showed that HSP27 is also involved in protein sumoylation, in the case of diabetes. It was found that HSP27 helps in increasing the number of cell proteins modified by small ubiquitin-like modifier (SUMO)-2/3. In stressed cells, HSP27 form large oligomer which binds to HSF1 and enters the nucleus which induces SUMO-2/3modification. Hence by this study we can say that HSP27 can be use as SUMO-E3 ligase specific for SUMO 2/3.

Virtual screening of ligands and then docking against SUMO protein was performed using PatchDock and Swissdocksoftwares. ZINC53683754 [(2S)-8-{(tert-butoxycarbonyl) amino}-2-(1H-indole-3-oyl)octanoic acid] andZINC53683750 [(2S)-2-(1H-indole-3-yl)hexanoic acid] comes out to be good SUMO protein binders.

This study proves that sumoylation has a role in the regulation of proteins involved in glucose metabolism. It identifies a new mechanism for the study of functions of SUMO proteins at the post-translational level and helps in identification of potential SUMO binding targets and potential ligand binders.

It was found that drug docking significantly reduces the number of experiments and thereby cut cost while examining the utility of any chemical as a drug before going through any in vivo or in vitro analysis.

INTRODUCTION

Diabetes Mellitus:

Diabetes mellitus, commonly called as diabetes. It is a metabolic diseases in which sugar level in blood increases to a dangerous level in a person. This may occur by any of two reasons 1) insulin cannot be produced by body. 2) Cells do not respond to the insulin that is produced. This can lead to serious health complications. High blood sugar can harm organs and raises the risk of heart disease.

Classification: Diabetes mellitus is classified into IDDM, NIDDM and Gestational.

IDDM (Type 1 diabetes) develops when insulin-producing beta cells of pancreas damaged or depleted by any reasons, because of which there occurs insulin deficiency in the body. The preventive measure against type 1 diabetes is not known and it is difficult to diagnose during early stages. Children or adults can be affected by it, but because a majority of these diabetes cases were in children, it was traditionally termed as "juvenile diabetes"

NIDDM (Type 2 diabetes) develops due to insulin resistance by the body. The insulin receptor believed to involve in responsiveness of body tissues towards insulin is defective. However, specific defects are not clearly known. Occurrence of type 2 diabetes is the most prevalent. During early stage, the predominant abnormality is reduced insulin sensitivity. The insulin sensitivity or reduce glucose production by the liver can be improve at this stage, by reversing hyperglycemia using variety of measures and medications.

GDM (Gestational diabetes mellitus) resembles type 2 diabetes in many ways. Occurrence of GDM is about 2%–5% of all pregnancies. After delivery it may improve or disappear. Gestational diabetes can be treated fully by careful medical supervision throughout the pregnancy. Type 2 diabetes may develop in about 20%–50% of affected women later in life.

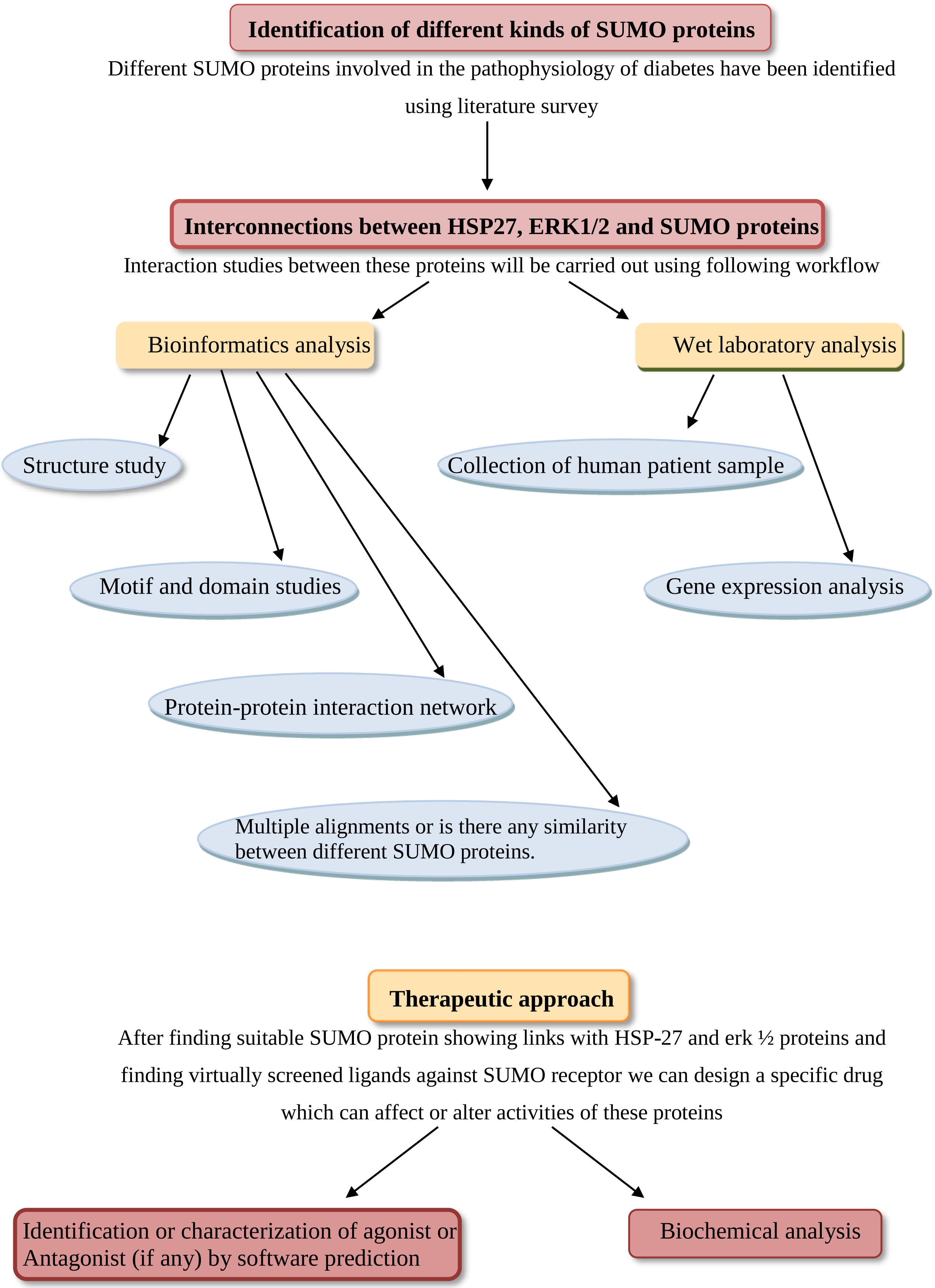
The focus of my study is Non-Insulin Dependent Diabetes Mellitus (NIIDM) which affects adults and contributes major proportion among diabetes patients worldwide.

Motivation for the project:

Despite the lots of drugs available for diabetes treatment, the disease remains a worldwide public health problem. Continuous efforts on the development of new drugs are required, and primary methods involve use of different approaches, such as testing natural products and synthetic molecules while advanced approaches involves study of physical and chemical properties of proteins, functional sites prediction, and finding out key protein-protein interactions in the biological processes which can be used for development or finding of efficient chemical compounds or drugs.

Proposed method:

Figure.1:



Identification of different kinds of SUMO proteins

Different SUMO proteins involved in the pathophysiology of diabetes have been identified using literature survey

Interconnections between HSP27, ERK1/2 and SUMO proteins

Interaction studies between these proteins will be carried out using following workflow

Bioinformatics analysis

Structure study

Motif and domain studies

Protein-protein interaction network

Multiple alignments or is there any similarity between different SUMO proteins.

Wet laboratory analysis

Collection of human patient sample

Gene expression analysis

Therapeutic approach

After finding suitable SUMO protein showing links with HSP-27 and erk ½ proteins and finding virtually screened ligands against SUMO receptor we can design a specific drug which can affect or alter activities of these proteins

Identification or characterization of agonist or Antagonist (if any) by software prediction

Biochemical analysis

Comparative docking studies were performed using two softwares namely PatchDock and SwissDock. FINDSITE tool was used to perform virtual screening and databases DrugBank and ZINC are used for getting structures of chemical constituents.

The strength of binding affinity between two molecules can be predicted with the help of prior knowledge of the preferred orientation. For the prediction of binding orientation, affinity and activity of small molecule to their protein targets, docking is frequently used.

Virtual screening of compound libraries is a standard tool in modern drug discovery. With the help of structure of target molecule, molecular docking can be used to discriminate between supposed binders and non-binders in large databases of chemical compounds. It can also be used in reducing the number of compounds for experimental testing. Visual examination of predicted docking poses (binding geometries) contributes in further development of a lead compound.

PyMOL software is a molecular graphics system which is used for viewing molecular structure. PyMOL have many 3D operations because of which it has many applications and programs which make it versatile for visualization of molecules.

Thesis Outline

The remaining part of the thesis is organized as follows. Review of literature describes the published literature concerning Type-2 diabetes and proteins under study. Methodology establishes the theoretical foundations and procedures behind the proposed technique and gives an overview of it. It also describes major steps of the proposed approach in details. The proposed approach was implemented under results section. Various analyses were performed on the obtained results. Finally, conclusions was summarised from this research, highlights the anticipated benefits and provides suggestions for future extensions of this work.

REVIEW OF LITERATURE

Noninsulin-dependent diabetes mellitus (NIDDM)

Diabetes mellitus type 2 (noninsulin-dependent diabetes mellitus (NIDDM) or adult-onset diabetes) is a metabolic disorder which happens when there is increase in blood glucose level with respect of insulin resistance while in diabetes mellitus type 1there is an absolute insulin deficiency due to destruction of islet cells in the pancreas.

Type 2 diabetes contributes majority of cases in all diabetes patients (90% of cases). It was thought that the primary cause of type 2 diabetes is obesity in people who are genetically predisposed to the disease.

Metformin or insulin may be needed as meditations if blood glucose levels are not adequately lowered by exercise and dietary medications.

Complications:

Long-term complications due to high blood sugar are-heart disease (two to four times the risk of cardiovascular disease), strokes, diabetic retinopathy (where eyesight is affected), kidney failure, and poor circulation of blood in limbs which results in amputations. The ketoacidosis (an acute complication) which is a feature of type 1 diabetes, is uncommon here.

Type 2 diabetes is the largest cause of non-traumatic blindness and kidney failure in developed countries.

It is associated with a ten-year-shorter life expectancy and increased rates of hospitalizations.

Other complications include: acanthosisnigricans, frequent infections, andsexual dysfunction.

Causes:

Type 2 diabetes is caused by a combination of lifestyle and genetic factors.

Some of the causes are controllable, such as diet and obesity, while causes, such as increasing age, genetic susceptibility, and female gender are not under control.

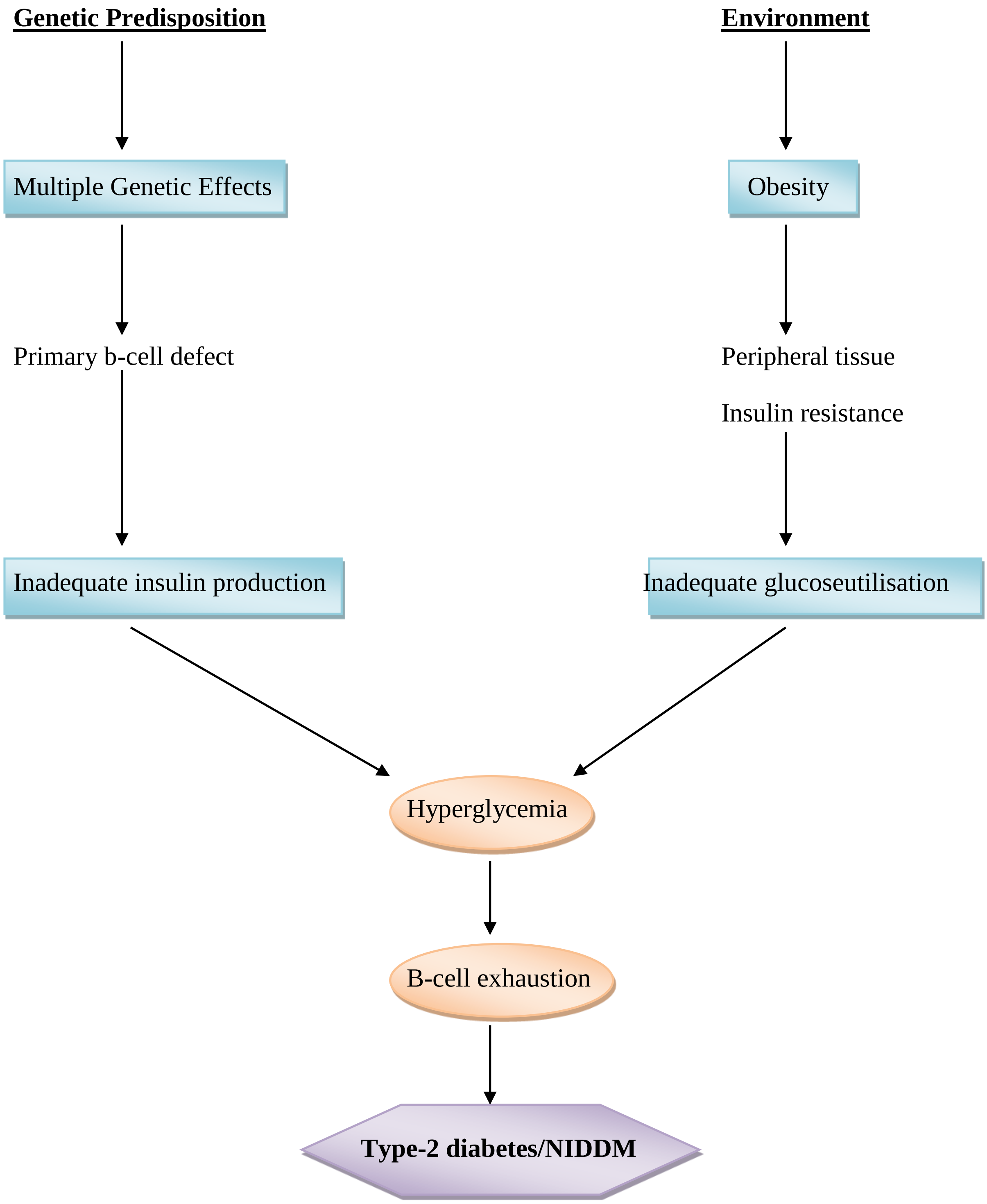


Figure.2: Pathogenesis of Type-2 *Diabetes mellitus*

THE CHAPERONING SYSTEM: PHYSIOLOGY AND PATHOLOGY

Chaperoning system:

The chaperoning system is a physiological set of molecules and molecular teams, and pertinent cells and tissues, which maintains protein homeostasis and other cellular functions.

Molecules, cells and tissues of the chaperoning system:

Chaperones are made in cells for work in the same cells or they can be exported to other cells. Chaperones for export are made in a cell and then travel to other locations, inside cells or in extracellular sites, in which they will take residence and work. It can be predicted that there are cells, and tissues or defined zones within certain tissues or organs, specialized in the production of chaperones for export.

The entire chaperone population of an organism is called chaperoning system. Each cell has its own set of chaperones or subsystem which includes more than one chaperoning complex or team, and teams interact forming networks inside the cell.

Classifications of molecular chaperones:

1. Types of chaperones according to their size (molecular weight):

Heat-shock proteins (HSP) are those proteins whose productions are induced whenever there occurs temperature elevation (heat shock), but the name is also applied to proteins from genes inducible by any other stress.

Many HSP are chaperones (e.g. HSP70 in humans) but not all chaperones are HSP (e.g. Alpha-Haemoglobin Stabilizing Protein). A.H.S.P is a dedicated chaperone whose substrate is alpha haemoglobin chain and is encoded by a gene which is not known. Conversely, many HSP are not chaperones (e.g. HSP32). HSP32 is associated with the generation of biliverdin and bilirubin, potent antioxidants.

Chaperone subpopulation		
Name	Other Names	MW (kDa)
Heavy	High MW, HSP 100	100 or higher
HSP90		81-99
HSP70	Chaperones, DnaX	65-80
HSP60	Chapronins, Cpn60	55-64
HSP40	DnaJ	35-54
Small HSP	sHSP, alpha-crytallins	34 or less
Other	Proteases: Isomerases: AAA+ proteins (e.g. paraplegin, spastinI) etc.	various

Table.1: Classification of HSP-chaperones according to molecular weight:

2. Types of chaperones according to their origin with regard to their place of residence and work:

Chaperones are found inside various cell compartments as well as in biological fluids moving around. For first case, the cells of origin and of residence are same, e.g. an autochthonous chaperone which resides and works in the same cell in which it was produced. As far as second case is considered, the place of residence (a given cell) of an imported chaperone is not the same as that of its origin.

Location	**Compartment**
Cellular	Nucleus Cytosol Mitochondria Endoplasmic Reticulum Lysosomes Vesicles Membrane on the inside Chloroplasts
Pericellular	Membrane on the outside
Extracellular	Intercellular space Blood (plasma, serum) Cerebrospinal fluid Secretions (e.g. saliva)

Table.2: Places at which chaperones reside and work

3. Types of chaperones according to their ability to move and change residence:

Chaperones can also be classified according to their mobility:

Sessile: fixed, anchored to another structure (e.g., cell membrane).

Mobile: not fixed, capable of moving inside a cell (e.g., from cytosol to nucleus), or outside cells (e.g., in the blood) and change place of residence and work.

Mobile chaperones can be of two subtypes:

Sedentary: reside always in the same cell or cell compartment but they can move within the region and

Nomadic: chaperones travel and work in various successive places. e.g. HSP60 is produced in the cytosol and then translocated to the mitochondria from which it can exit and go back to the cytosol, and even exit the cell and appear in the extracellular space.

4. Types of chaperones according to their relation with other chaperones or other molecules:

Chaperones exercise their functions alone or in associations with other molecules. They can be considered

Single: one chaperone molecule performs its role and is not associated with any other chaperone.

Social: chaperones form part of a Team (Chaperoning Team- a specific association of chaperones to build a chaperone machine).

Network: (Chaperoning Network), which is a specific interaction between chaperone machines (e.g., HSP70-HSP40- NEF, and HSP60-HSP10, and Prefolding), or between a chaperoning team and a single chaperone.

Heat Shock Protein-27 (HSP-27):

HSP27 is a chaperone which belongs to sHSP (small heat shock protein) group (12–43kDa).

The common functions of small HSPs are: thermotolerance, chaperone activity,inhibition of apoptosis, signal transduction,regulation of cell development, and cell differentiation.

Proteins in small HSP family including HSP27 share α-crystallin domain, a conserved c-terminal domain. Initially HSP27 was marked as a protein chaperone which helps in proper refolding of damaged proteins. Later it was found that HSP27 protein responds to cellular stress conditions like oxidative stress and chemical stressother than heat shock. HSP27 functions as an antioxidant during oxidative stress, in which it lowers down the levels of reactive oxygen species (ROS) by increasing the levels of intracellular glutathione and lowering the levels of intracellular iron.

Under chemical stress HSP27 act as an anti-apoptotic agent and interact with both mitochondrial dependent as well as independent pathways of apoptosis. HSP27 also provides protection from programmed cell death by inhibition of caspase-dependent apoptosis.

During heat shock and other stress conditions, HSP27 can also regulate the dynamics of actin cytoskeleton, which stimulates polymerization and capping of actin protein by itself.

Oligomerization:

The oligomerization of HSP27 inside cellular environment is a dynamic process. There is a balance between stable dimers/tetramers and instable oligomers consisting of 16 to 32 subunits

This process depends on the phosphorylation status of HSP27, physiology of the cells, and the cell exposure to stress. Due to stress there occurs an increase of phosphorylation of HSP27 after several minutes and their expression after hours. The formation of oligomers is due to phosphorylation of HSP27.

The chaperone activity is because of its oligomerisation. The higher chaperone activity is due to large oligomers, and by dimers which have no chaperone activity. Therefore under heat shock, formation of large aggregates takes place.

Functional regulation by phosphorylation:

The levels of HSP27 is lowest in cells and tissues. HSP27 is organized as large oligomers. When P38 MAPK pathway is activated at multiple serine residues (15, 78, and 82) in protein, HSP27 is phosphorylated by MAPKAP protein kinase 2/3.

After phosphorylation, HSP27 converted into smaller oligomers, often dimers and tetramers which can interact with other proteins.

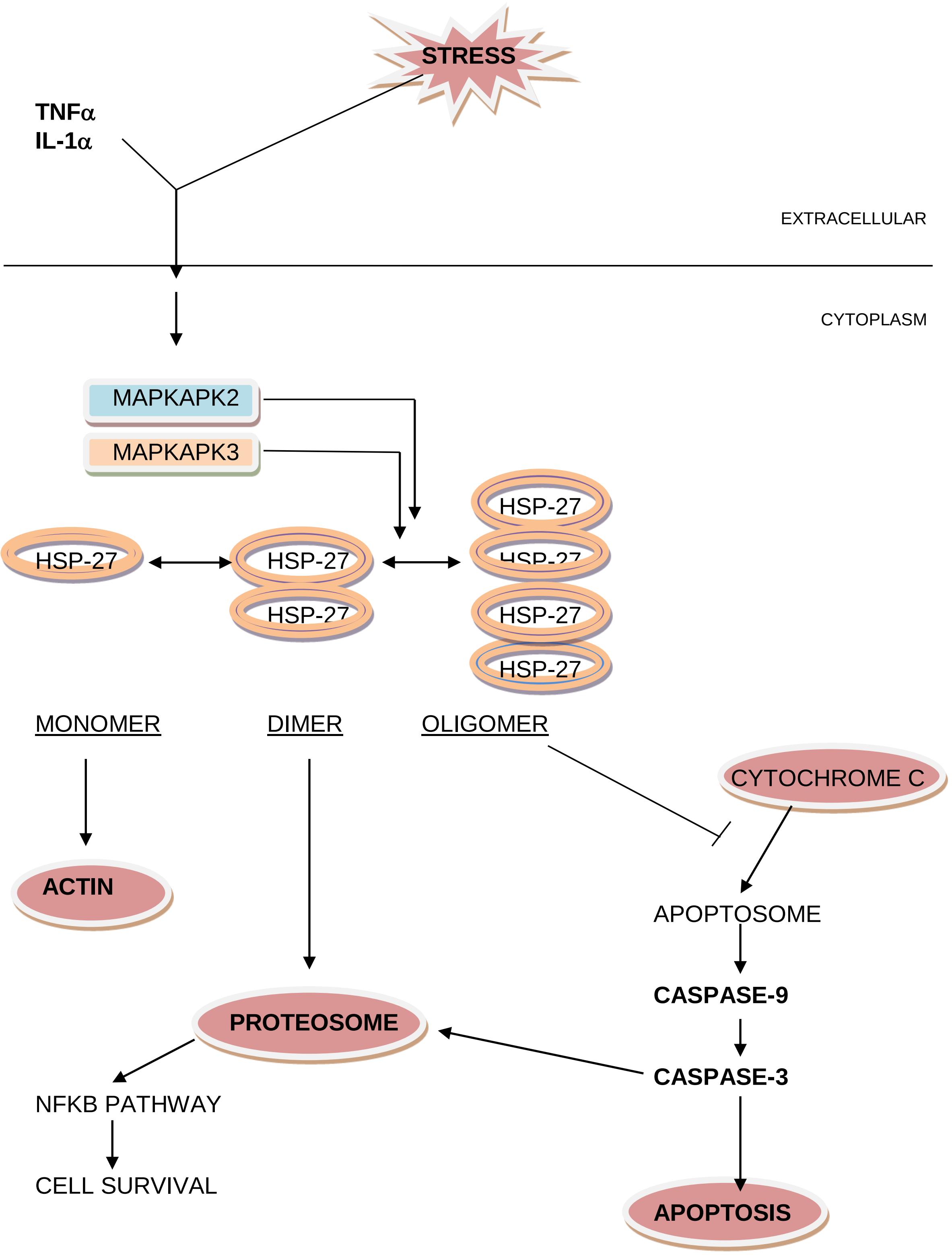


Figure.3: Functional regulation of HSP-27

Extracellular-signal-regulated kinases (ERKs)

Extracellular signal regulated kinases (ERKs) are protein kinase intracellular signalling molecules which are widely expressed in cell and tissues.

ERKs regulates meiosis, mitosis, and post-mitotic functions in differentiated cells.

Many factors like growth factors, ligands forheterotrimeric G protein-coupled receptors, cytokines, transforming agents, virus infectionand carcinogens stimulates ERKs productions.

MAPK/ERK pathway:

The term "extracellular-signal-regulated kinases", is sometimes refered as mitogen-activated protein kinase (MAPK). In the MAPK/ERK pathway, c-Raf isactivated by Ras, which then followed by mitogen-activated protein kinase (abbreviated as MAP2K, MKK or MEK) and then MAPK1/2. Growth hormonesreceptor typically activates Ras through tyrosine kinases and GRB2/SOS, which may also receive other signals. ERKs can activates many transcription factors, like ELK1 and some downstream protein kinases. In cancersERK pathway may bedisrupted commonly, especially Ras, c-Raf and receptors such as HER2.

Types of ERKs:

1. Mitogen-activated protein kinase 1 (MAPK1):

The other known name of Mitogen-activated protein kinase 1 (MAPK1) is "extracellular signal-regulated kinase 2" (ERK2). They were found while searching protein kinases which are rapidly phosphorylated after activation of cell surface tyrosine kinases like the epidermal growth factor receptor. ERKs phosphorylation results in activation of their kinase activity.

Ras GTP-binding proteins were found to be involved in the activation of ERKs. Another protein kinase, Raf-1, can phosphorylate a "MAPK kinase", thus called as a "MAPK kinase kinase". The MAPK kinase was named "MAPK/ERK kinase" (MEK).

Receptor-linked tyrosine kinases, Ras,Raf, MEK, and MAPK are involved in signalling cascade linking an extracellular signal to MAPK activation.

2. Mitogen-activated protein kinase 3 (MAPK3):

The other known name of Mitogen-activated protein kinase 3 (MAPK3) is "extracellular signal-regulated kinase 1" (ERK1). It is thought that MAPK1 are capable of fulfilling most of the MAPK3 functions in many cells but the main exception is in T cells. There is reduced T cell development after $CD4^+CD8^+$ stage in mice lacking MAPK3.

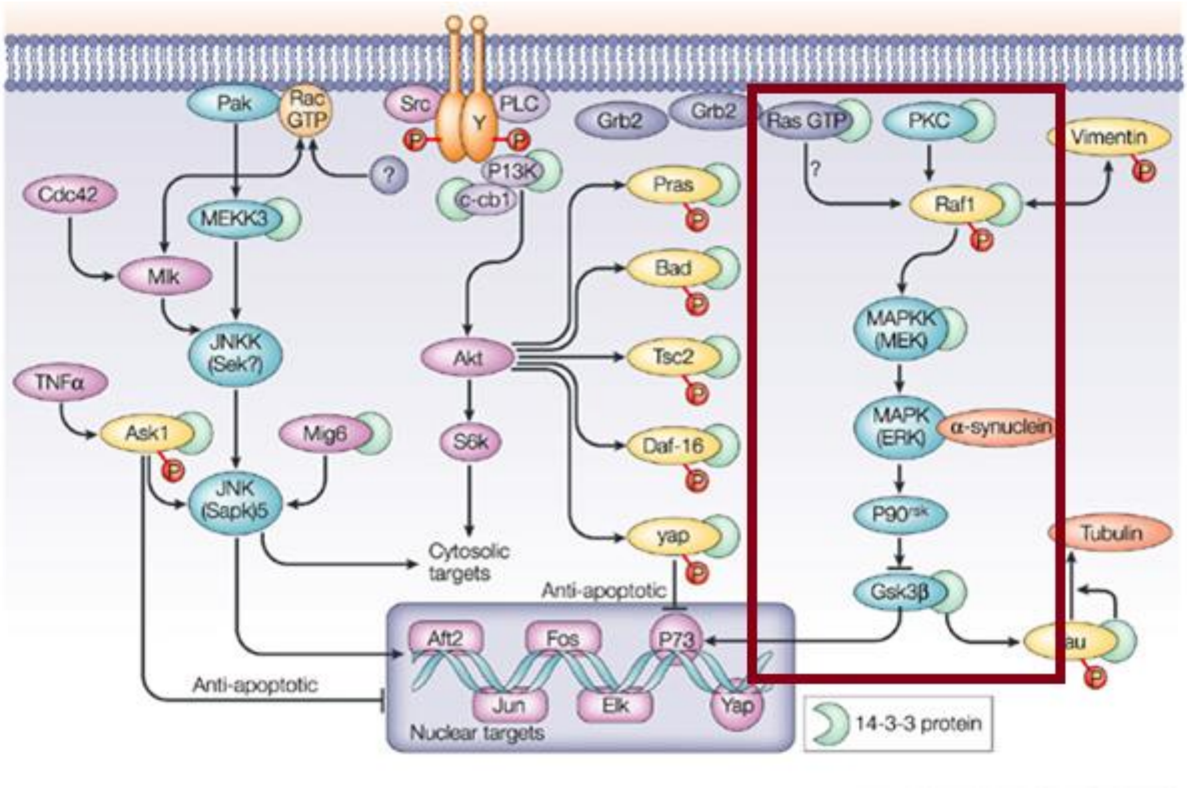


Figure.4: 14-3-3 Proteins (green) are primarily involved in the serine/threonine protein kinase Akt pathway and in the extracellular signal-related kinase pathway (ERK1/2) *(Source: Nature reviews: General role of 14-3-3 proteins in the mitogen-activated protein kinase (MAPK) pathway)*

SUMO proteins:

Small Ubiquitin-like Modifier (SUMO) proteins belongs to family of small proteins that can attach or detach covalently with other proteins to modify their function in cells.

SUMOylation is a kind of post-translational modification of proteins. It is involved in various cellular processes, stress response, nuclear cytosolic transport, protein stability, transcriptional regulation and apoptosis.

SUMO proteins are similar to ubiquitin protein. The SUMOylation process is directed by same enzymes which are involved in ubiquitination. Unlike ubiquitin, SUMO proteins are not used to tag proteins for degradation.

Four isoforms of SUMO proteins are found in humans; SUMO-1, SUMO-2, SUMO-3 and SUMO-4. SUMO2 and SUMO3 show a high degree of similarity to each other (97% identity) hence they are highly homologous protein and are distinct from SUMO-1 protein.

Sumoylation regulates the molecular interactions of the modified proteins are regulated by SUMOylation which leads to changes in the activity, solubility, localization, or even stability of respective target proteins.

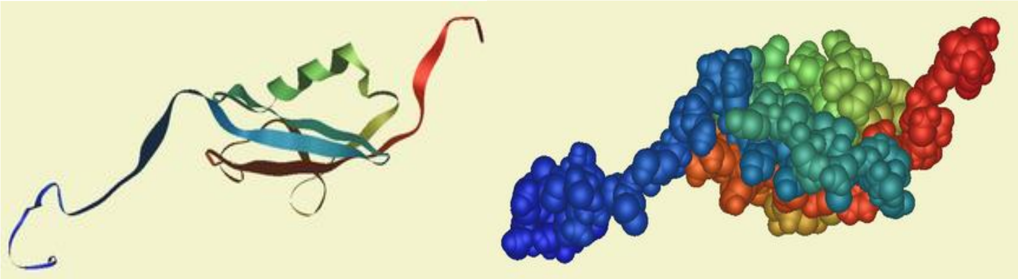
SUMO isoforms regulates different mitotic processes in mammalian cells. In the cell cycle during mitosis stage, SUMO-2/3 get localise to centromeres and condensed chromosomes, whereas SUMO-1 get localise to the mitotic spindle apparatus. During mitosis SUMO-2/3 conjugate with topoisomerase II and exclusively modified it. This is one of the major SUMO conjugation products associated with mitotic chromosomes.

During stress condition SUMO-2/3 modifications seem to be take place. SUMO-1 and SUMO-2/3 can form mixed chains, however, SUMO-1 is thought to terminate these poly-SUMO chains because it does not contain the internal SUMO consensus sites found in SUMO-2/3. SUMO-1 is sometime called as 'modified modifier' because its Serine 2 is phosphorylated.

Structure:

SUMO proteins globular and smaller proteins with both ends of the amino acid chain coming out of the protein's centre. An alpha helix and a beta sheet is present in its spherical core.

Most of SUMO proteins are around 100 amino acids long and have 12 kDa as mass. Different SUMO family members have different length and mass of proteins. This variation is also because of proteins belonging different species or organism. SUMO has a nearly identical structural fold with ubiquitin because of its little sequence identity at the amino acid level.



Structure of SUMO protein *(Source: PDB)*

Functions:

The main function of SUMO is post-translational modification on lysine residues of proteins. This modification plays a crucial role in a number of cellular processes and is done by covalent attachment via an isopeptide bond.

SUMO modification of proteins has many functions e.g. nuclear-cytosolic transport, protein stability and transcriptional regulation. SUMOylation occurs on very small fraction of a given protein as a result of which a number of different outcomes (altered localization) takes place. The first identified SUMO substrate is RanGAP1. The SUMO-1 modification of RanGAP1 leads to its movement from cytosol to nuclear pore complex. Similarly the SUMO modification of hNinein results to its movement from centrosome to the nucleus. Also it was observed that the inhibition of transcription generally correlates with SUMO modification of transcriptional regulators in most of the cases.

Figure.5: Benefits of SUMO platform

Benefits of SUMO platform

Problem	**Solution**
Low yield	SUMO-enhanced expression
Specific N-terminus required	SUMO specific processing
Insolubility	SUMO-driven folding
High costs of goods	SUMO affinity production

METHODOLOGY

1. Protein-Protein Interaction:

With the help of genomic associations between the genes, functional links between their proteins can be found. Genes that are required for a particular function are found in similar species, tend to be involved in gene-fusion events and are located in close proximity on the genome (in prokaryotes).

STRING database:

The STRING database is used for the inspection and analysis of associations between genes. A unique scoring-framework is found in STRING which is based on different types of association standards which is integrated in a single confidence score per prediction.

To identify or predict functional associations for a protein of interest STRING database is used. Every protein have a unique accession number or identifier so we can start search by using accession no. of given protein. The other way for identifying the corresponding entry in the database is by supplying the raw amino acid sequence of the protein for checksum lookups and similarity searches. The functional links for the protein is then predicted which are ranked by their estimated confidence.

Data Sources:

The experimentally derived protein-protein interactions data through literature curation is imported in STRING like many other databases that store protein association knowledge. The computationally predicted interactions are also stored in STRING from: (i) scientific texts minings, (ii) computation of interactions from genomic features, and (iii) orthology based interactions transferred from model organisms.

The standardisation of all predicted or imported interactions are done against a common reference of functional partnership as annotated done by KEGG (Kyoto Encyclopaedia of Genes and Genomes).

1. Imported data

The knowledge of protein association from physical interaction databases and biological pathway databases (MINT, KEGG, HPRD, BIND, Reactome, DIP, BioGRID, NCI-Database, GO,IntAct, EcoCyc) are imported in STRING.

2. Text mining

To search for co-occurrences of gene names which are statistically relevant, lots of scientific texts (OMIM, SGD,PubMed, and FlyBase) are parsed.

3. Data prediction

Neighbourhood: Similar function of the proteins in different species is suggested by similar genomic context.

Fusion-fission events: Proteins that are likely to be fused in some genomes if they are functionally linked.

Occurrence: In a metabolic pathway proteins having similar function must be expressed together. Phylogenetic profiles of such proteins also similar.

Co-expression: The predicted association between genes leads to simultaneous expression of genes whose patterns are observed.

VisANT: Shortest Pathway analysis

Biomolecular interaction data is integrated by VisANT into a graphical interface. Data flexibility is arranged in layers by this software for fast retrieval of data in this visualization and analysis package.

VisANT is an online tool which is freely available. A large range of data sets about biomolecular interactions are provided by online interface. This system is integrated with GenBank, KEGG and SwissProt and other standard databases for annotation.

For studying wide range of biological applications like gene regulation, pathways and systems biology, VisANT is used.

For mining and visualizing, pathway, sequence, structure and associated annotations VisANT software is efficiently used. Due to presence of a variety of built-in functions, data about predicted association and interaction can be manipulated, combined, analysed and overlaid.

Searching of node by VisANT:

Firstly the node is search in the existing network to check whether it is present there or not. If node is found then it is selected. If it is not found then VisANT server will try to find it in the dictionary of Predictome database. Interaction data will be returned to VisANT if node is found in the dictionary. Hence expanded linkages are sometimes found in the search. On the screen it sometime just showed selected nodes in the network. A warning window will be displayed if no entry is found for the searched object.

2. Physico-chemical properties:

PROTPARAM

ProtParam is the protein analysis tool. This tool is available freely on the ExPasy server. Various physiochemical parameters of a protein can be calculated by using it. Several parameters are calculated by using the protein sequence as its input. SwissProt or TrEMBL accession number of protein can be used as input on the server.

ProtParam calculates various parameters such as extinction coefficient, molecular weight instability index, aliphatic index, amino acid composition, grand average of hydropathicity, estimated half-life, and theoretical pI.

Molecular Weight

In ProtParam, the addition of average value of isotopic masses of amino acids in protein helps in calculation of molecular weight of protein.

Extinction coefficient

The extinction coefficient tells about amount of absorption of certain wavelength of light by a protein. The molar extinction coefficient of a protein can be calculated from the information of its amino acid composition. The extinction coefficient of a given protein can be calculated with the help of molar extinction coefficient of the tyrosine, tryptophan and cystine protein. This can be calculated by using the equation given below:

E (Prot) = Numb (Tyr)*Ext (Tyr) + Numb (Trp)*Ext (Trp) + Numb (Cystine)*Ext (Cystine)

Calculation of optical density or absorbance can be done by using the formula given below:

Absorbance = E / Molecular weight

Theoretical pI

Protein pI can be calculated by using pKa values of amino acids. Due to variations in side chains, pKa value of Amino acids also varies. For defining the pH dependent characteristics of a protein protein pI value is used.

Half-life

The half-life is a time taken to disappear half of the amount of protein in a cell after its production.

Grand average of hydropathicity (GRAVY)

By adding and dividing the hydropathy values of each amino acids by the number of residues in the sequence helps in calculation of GRAVY value for a protein or a peptide. Increasing positive score indicates greater hydrophobicity.

Aliphatic index

The aliphatic index of a protein is the relative volume occupied by the amino acids having an aliphatic side chain in their structure like alanine, valine, isoleucine and leucine. The aliphatic index of a protein can be calculated by the following formula.

Aliphatic index = X(Ala) + a * X(Val) + b * (X(Ile) + X(Leu))

Instability index

It tells about the estimate of the stability of a protein in a given sample. If instability index is smaller than 40 then protein is stable, and if it is above 40 then protein may be unstable.

3. Protein variations:

SwissVar

SwissVar is an online server for searching variants in the UniProt Knowledgebase (UniProtKB) having Swiss-Prot entries. Swiss-Prot Variant pages are accessed directly by it.

All the information about a particular variant are summarized in Swiss-Prot Variant pages. These pages contain:

- Annotation which is done manually are based on literature. For each specific variant these annotations tells genotype-phenotype relationship.
- Conservation scores and structural features helps in assessing the effect of the variant.

Three types of searches can be performed by using SwissVar:

Protein: Search is done using, UniProt accession number or simply by protein or gene names

Functional/structural features: Search is done using functional and structural parameters of the variant.

Disease: Search is done using OMIM identifier, MeSH terms and also simply by disease names.

4. Multiple Sequence Alignments:

Multiple sequence alignments are used for analysing many sequence together. Progressive heuristic alignment method is used for computation. Multiple sequence alignments involve comparing homologous sequences which is essential in most of bioinformatics analyses.

CLUSTAL OMEGA

Clustal Omega is a program based on multiple sequence alignment between proteins. Biologically meaningful multiple sequence alignments are produced for divergent sequences. The best match for the selected sequences is calculated by this program. It then lines them according to the similarities and differences between sequences. With the help of Cladograms or Phylograms Evolutionary relationships can be viewed.

Global alignment is done for sequences by aligning their entire length while local alignment is done for certain specific regions. This is true for pair-wise and multiple alignments. Gaps representing insertions/deletions are used in global alignments while same are avoided in local alignments. The alignment is progressive and hence considers the sequence redundancy. From multiple alignments, trees can also be calculated.

At the bottom of the submission form, all sequences in FASTA format are copied into the open frame. It must be ensure that one blank space must be given between all sequences. Based on their similarities, these amino acid sequences are then aligned by Clustal Omega. After running the program results appears after a few seconds.

The alignment score and the number of sequences submitted are shown in the Scores Table along with other information.

Steps involved:

- In first step user provided input (e.g. sequences, databases.)
- In the following steps, the default tool parameters can be change or altered according to the needs of user.
- In the last step which is tool submission step. A title is specified by the user with which result is associated. Information will be submitted by using submitting button in the program. An email notification will be send at provided email address when computation is finished.

5. SUMO binding sites:

SUMOsp

GPS and MotifX are the two methods used by SUMOsp for prediction of SUMOylation Sites. Prediction of phosphorylation site is done originally by GPS and MotifX. These two pattern recognition methods are accurate and robust for the prediction of sumoylation site as there is 5-fold cross validation.

For scoring potential sumoylation peptide sites PSSM algo is used. Sites which are followed by ψKXE motif are predicted. Unusual sumoylation sites are also predicted which are present in other non-canonical motifs

A set of highly-specific motifs for the sumoylation sites is generated by MotifX. Some of these sumoylation sites are, LKXE, KXE, VKXE, IKXEP and IKXE where X can be any amino acid. On combination with GPS, MotifX exhibits greater computing power.

6. Domain and Motif:

PROSITE

The function of an uncharacterized proteins can be determined with the help of PROSITE. Biologically significant sites and patterns are present in this database. It can reliably and rapidly identify family of protein with its computational tools.

For determining the functions of proteins by using protein sequence patterns is becoming one of the neccessary tools of sequence analysis.

Short sequences can find certain active sites or binding properties. These short sequences can be grouped into a small subgroups and then searched against your sequence.

In some cases, for an unknown protein which is related distantly with any known protein, the structure and function of this unknown protein can be detected by whole sequence alignment for the seach of motifs, a particular cluster of residues. The motifs, fingerprints or templates lay very tight restraint on the evolution of regions of a protein sequence. These motifs are important specific regions in a protein, because of which they have enzymatic activity or binding properties.

7. Binding sites in protein:

3DLigandSite

3DLigandSite is an online server for predicting ligand-binding sites. Protein-structure prediction is utilised by 3DLigandSite for providing unresolved proteins structural models. Ligands bound to structures are superimposed onto the model if they are similar to the query and then used for predicting the binding site. In homologous structures, protein structures and the ligands are founded by 3DLigandSite for predicting ligand binding sites.

Method:

1. Prediction of structures

After submission of sequence, the protein structure is predicted using the Phyre server. If the user has provided their own structure then this step is not required.

2. Structural Library of Ligand bound structures searching

Structures which are homologous to the query and have bounded ligands are identified. For performing a full structural scan of the modelled structure, MAMMOTH algo is used. This searching is done against a library of protein structures having bound ligands. For analysis, top 25 scoring ligands are retained.

3. Clustering of Ligand & prediction of binding site

For grouping ligands, single linkage clustering is used. The selection of the cluster having most number of ligands takes place and residues are predicted for forming the binding site which are within a threshold distance of the clustered ligands. The Jensen Shannon divergence score is used for residue conservation. This score is then provided as a guide for the user.

8. Virtual screening:

Virtual screening of ligands is a widely used process in assisting new pharmaceutical discovery. The development of new methodologies is required as virtual screening approaches have many limitations.

FINDSITE:

FINDSITE is a threading-based virtual screening process. Structural information which are extracted from weakly related proteins are employed in it. This information can be used to perform rapid ligand docking and on the basis of homology modelling of protein structures ranking is done. FINDSITE uses all-atom ligand docking approaches for low-quality modelled receptor structures which increases the accuracy of ligand binding pose prediction and therefore requires less CPU time.

The target ligand are superimposes onto the consensus binding pose using FINDSITE docking. The anchor substructure is identified when conformation of anchor averaged over the seed compounds (4 Å RMSD).

The superposition of multiple conformations of the target ligand is done due to ligand flexibility. The conformation having lowest RMSD value can be superposed onto the reference coordinates. The final model is produced by selection of predicted pose.

Information about chemical properties of the binding ligands are also provided by FINDSITE. Then representative ligand molecules are selected using this observation. These ligands are more likely to bind on the predicted site on surface of protein. Subsequently, in a simple ligand-based virtual screening experiment, ligand templates were used against the KEGG compound library.

Scoring functions for virtual screening:

For ligand-based virtual screening, FINDSITE uses motif-based method for ranking the screening library. The ligands which are weakly homologous or having <35% sequence identity to the target are identified by PROSPECTOR 3 algo having Z-score value ≥4. For motif-based virtual screening, ligand templates are constructed by FINDSITE which cluster the molecules and simultaneously rank them.

9. Protein- Ligand docking:

PatchDock:

PatchDock is used for molecular docking of protein structure with ligand. Two molecules of any type can be used as input: drugs, DNA, peptides, proteins. A list of potential protein complexes showing complementarity in shape are presented in output.

Let us take two molecules, the surfaces of these molecules are fragmentised into different patches due to different shape of surface. Patterns are produced by using these patches. Once, they can be superimposed, the patches are identified using shape matching algorithms.

The three major stages in PatchDock algorithm are:

1. Representation of molecular Shape - the molecular surface of the molecule is computed in this step. After this, for detection of geometric patches (flat, convex and concave surface pieces) a segmentation algorithm is applied. Only patches having 'hot spot' amino acids are retained after filtering.

2. Matching of Surface Patch - the Geometric Hashing technique and Pose-Clustering matching technique are applied for matching the patches detected in the previous step. Concave patches are matched with flat and convex patches.
3. Filtering and Scoring – in this step examination of the candidate complexes takes place. The penetrations of the atoms of the receptor in to the atoms of the ligand are not accepted and complexes showing them are discarded. Finally, the ranking of remaining candidates takes place according to a geometric shape complementarity score.

There are different sets of parameters in PatchDock, which are optimized for different types of complexes. The search space is restricted to the cavities of the enzyme molecule by the algorithm in enzyme-inhibitor type complex while in antibody-antigen complex, the searching is restricted to CDRs of the antibody detected by the algorithm (the antibody should be used as 'receptor molecule'). The algorithm uses parameter set optimized for small size molecules in case of protein-small ligand docking.

SwissDock:

It is a docking tool which is used for examination and prediction of interactions between molecules. For an example interaction between a small molecule and a target protein.

Its algorithm consists of the following steps:

1. Binding Modes (typically from 5000 to 15 000) are generated in large numbers by local docking where search is done in a user-defined box or by blind docking where search is done in the target cavities of the entire protein surface.
2. With the help of a grid, CHARMM energies of Binding modes are estimated.
3. Then, ranking of Binding Modes takes place according to most favourable energies.
4. Finally, result file contains the most favourable clusters.

Hence accurate docking assays is carried out by this unique combination of features within minutes.

RESULTS

1. Comparative study of SUMO proteins:

SUMO proteins are studied first on preliminary basis by performing literature surveys and collecting information from various online servers. It was found that SUMO proteins play important role in biological process and have important functions. The various physical and chemical properties and other features of SUMO proteins were collected in this study are shown below in table 3.

These parameters includes gene names, number of amino acid sequence, functions and biological processes, tissue specificity and post translational modifications.

Most of these properties were found to be common among all four SUMO proteins because of high sequence and structure homology between them.

All of these properties related to SUMO proteins are listed in the tabular form.

	SUMO-1	SUMO-2	SUMO-3	SUMO-4
Protein Names	SUMO-1, SMT3 homolog 3, GMP1, Sentrin, SMT3C, UBL1, PIC1	SUMO-2, SMT3 homolog 2, Sentrin-2,SMT3A, HSMT3	SUMO-3, SMT3 homolog 1, SMT3B	SUMO-4
Gene Names	SUMO-1,SMT3C, SMT3H3, UBL1	SUMO-2, SMT3A, SMT3H2	SUMO-3, SMT3B, SMT3H1	SUMO-4, SMT3H4
Seq. length (A.A.)	101	95	103	95
Functions	SUMO1 is involved in Signal transduction Nuclear transport, mitosis, DNA replication and repair. RANGAP1 is targeted by SUMO1 which lead to its movement towards the nuclear pore complex protein RANBP2. Proteasomal degradation of modified proteins occurs when SUMO1 chains bounded with protein are polyubiquitinised.	SUMO2 is involved in Signal transduction Nuclear transport, mitosis, DNA replication and repair. Proteasomal degradation of modified proteins occurs when SUMO2 chains bounded with protein are polyubiquitinised.	SUMO3 is involved in Signal transduction Nuclear transport, mitosis, DNA replication and repair. It may not be involved in protein degradation. It may act as an antagonist of ubiquitin during degradation process.	SUMO4 helps in Modulation of protein sub-cellular localization, activity or stability. It may not be involved in protein degradation During oxidative stress, SUMO4 form conjugates with various anti-oxidant enzymes, chaperones, and stress defence proteins.

PTM	Isopeptide bond,Ubl conjugation, Acetylation, Phosphoprotein	Isopeptide bond, Ubl conjugation	Isopeptide bond, Ubl conjugation	Isopeptide bond
Tissue specificity	Expressed broadly	Expressed broadly	Expressed predominantly in liver	Expressed mainly in adult and embryonic kidney and also in immune tissues, like lymph node and spleen.
Cellular component	-PML body -cytoplasm -dendrite -nuclear membrane -nuclear pore -nuclear speck -synapse	-PML body	-cytoplasm -kinetochore -nucleus	-nucleus
Biological Processes	Protein sumoylation Ubiquitin dependent protein catabolic process is positively regulated. Involved in interferon-gamma-mediated signalling pathway regulation and localization of protein at nuclear pore Sequence-specific DNA binding transcription factor activity is negatively regulated Assembly of protein complex ispositively regulated	Protein sumoylation Ubiquitin dependent protein catabolic process is positively regulated.	Protein sumoylation	Protein sumoylation

Table.3: comparative study of SUMO proteins

2. Variations among SUMO proteins:

Variations among different SUMO isoforms were studied with the help of SwissVar portal.

SUMO 1 variant

Disease/Polymorphism: Non-syndromicorofacial cleft 10

Cytogenetic location: 2q33.1

Variants Information:

Natural Variations: Alternative sequence 4-28. Missing in isoform 2

Experimental info:

Mutagenesis-

36. F → A: Abolishes binding to PIAS2.

Sequence conflict-

75. H → N in AAH66306.

The haplo insufficiency of SUMO1 was identified in a patient who is associated with isolated unilateral CLP. Strong expression of SUMO1 occurs in the upper lip, primary and secondary palate (medial edge epithelium) in the mouse. In human also a micro deletion surrounding SUMO1 expressing gene supports clefting in humans.

SUMO2 variant:

Disease/ Polymorphism: p.Asp16Asn

Cytogenetic location:17q25.1

Variants information:

Natural Variation:

Variant position: 16

Location on the sequence:

MADEKPKEGVKTENN D HINLKVAGQDGSVVQFKIKR

Type of variant: Polymorphism

Residue change: From Aspartate (D) to Asparagine (N)

Physico-chemical properties: Change from medium size and acidic (D) to medium size and polar (N)

Experimental Info:

Mutagenesis- 11 K → R: Abolishes the formation of poly(SUMO) chains

SUMO3 Variant:

Disease/Polymorphism: p.Pro38Ser

Cytogenetic location: 21q22.3

Variant information:

Natural Variation:

Variant position: 38

Location on the sequence:

NLKVAGQDGSVVQFKIKRHT P LSKLMKAYCERQGLSMRQIR

Type of variant: Polymorphism

Residue change: From Proline (P) to Serine (S)

Physico-chemical properties: Change from medium size and hydrophobic (P) to small size and polar (S)

Experimental Info:

Mutagenesis-

a) 11. K → R: Abolishes the formation of poly (SUMO) chains

b) 33. I → A: Impaired interaction with USP25; when associated with A-34.

c) 34. K → A: Impaired interaction with USP25; when associated with A-33.

Sequence conflict-

a) 32. K → E in BAD96311

b) 76. E → R in CAA67896

SUMO 4 Variant:

Disease/ Polymorphism: p.Met55Val

Cytogenetic location: 6q25.1

Variant information:

Natural Variation:

Variant position: 55

Location on the sequence:

KRQTPLSKLMKAYCEPRGLS M KQIRFRFGGQPISGTDKPAQ

Type of variant: Polymorphism

Residue change: From Methionine (M) to Valine (V)

Physico-chemical properties: Similar physico-chemical property. Both residues are medium size and hydrophobic.

Variant description: It may be associated with susceptibility to type 1 diabetes; IL12B expression and greater NFKB1 transcriptional activity. Variant Val-55 could be associated with insulin-dependent diabetes mellitus 5 (IDDM5).

3. Protein-Protein interaction network:

In the Protein-Protein interaction network study of SUMO proteins, the following results were obtained. It shows closely interacted proteins with SUMO proteins. All these interacted functional analysis is collected from String database itself along with the interacted proteins.

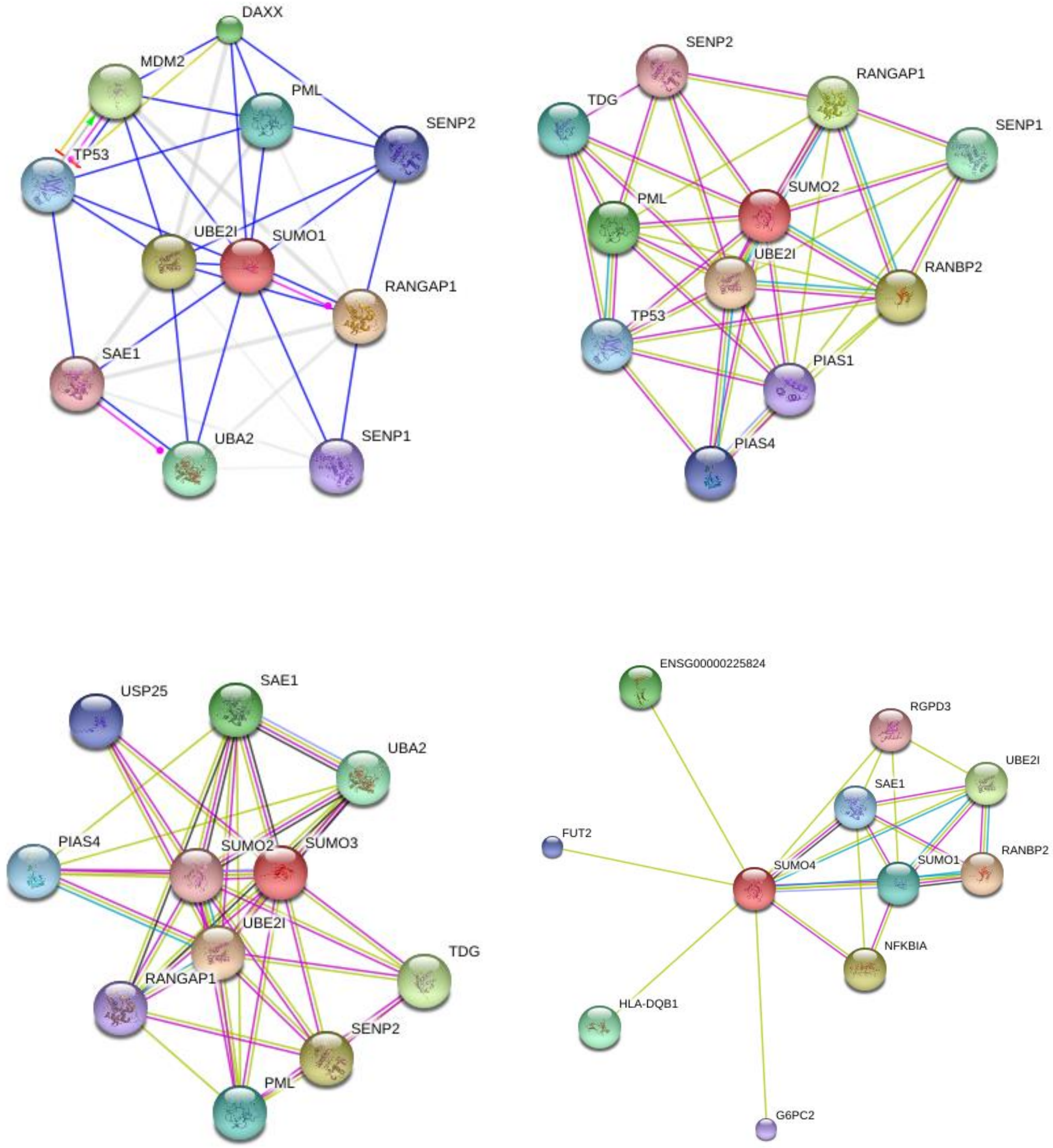


Figure.6: Protein-protein interaction network of SUMO proteins

SUMO proteins: they are ubiquitin-like proteins which are attached covalently to lysines of target proteins. They are attached as a monomer in case of SUMO1 and SUMO4 and as a lysine-linked polymer in case of SUMO2 and SUMO3. As these proteins does not seem to be involved in protein degradation, hence, they may act as an antagonist of ubiquitin in the degradation process. In many cellular processe,s they plays role like DNA replication and repair, nuclear transport, signal transduction and mitosis. The E1 complex SAE1- SAE2 requires for prior activation for attaching SUMO proteins covalently to their substrates and linkage to the E2 enzyme UBE2I.

RANGAP1- Ran GTPase activating protein 1; the nuclear Ras-related regulatory protein Ran is converted in to the inactive GDP-bound state with the help of this GTPase activator. RANGAP1 is targeted by SUMO1 to the nuclear pore complex protein RANBP2.

RANBP2- RAN binding protein2; SUMO1 and SUMO2 conjugation by UBE2 is facilitated by this protein. RANBP2 may also have chaperone or isomerase activity and may bind DNA or RNA. In nuclear export pathway RANBP2 is actively involved.

MDM2- It binds with transcriptional activation domain of p53 and p73 which leads to inhibition of cell cycle arrest and apoptosis. In the presence of E1 and E2, MDM2 act as ubiquitin ligase E3, against p53 and itself. It targets p53 for nuclear export and proteasome-mediated proteolysis.

DAXX- death-domain associated protein. Many proteins are interacted by this protein, like centromere protein C, apoptosis antigen Fas, and transcription factor erythroblastosis virus E26 oncogene homolog 1.

UBA2- ubiquitin-like modifier activating enzyme 2; this enzyme have dimeric structure and acts like an E1 ligase for SUMO1, SUMO2, SUMO3, and SUMO4. The activation of SUMO proteins and thioester bond formation with a conserved cysteine residue on SAE2 is mediated by it.

TP53- tumour protein p53; in many tumour types, they acts as a tumour suppressor. Depending on the physiological circumstances and cell type, it leads to induction of apoptosis or growth arrest. It plays a role in regulation of cell cycle by acting as a trans-activator due to which cell division is regulated negatively by controlling a set of genes. An inhibitor of cyclin-dependent kinases is also one of such activated genes. Stimulation of BAX and FAS antigen expression mediates induction of apoptosis. Repression of Bcl-2 expression also mediates apotosis.

SENP2- SUMO1/sentrin/SMT3 specific peptidase 2; two essential functions in the SUMO pathway are catalysed by this protease: SUMO1, SUMO2 and SUMO3 are processed to their mature forms and with the help of this protein SUMO1, SUMO2 and SUMO3 get deconjugated from targeted proteins. It may modulate the Wnt pathway by down-regulating CTNNB1 levels.

PML- promyelocytic leukaemia: the activity of the tetrameric form of PKM2 is inhibited when ELF4 is recruited into PML nuclear bodies. This protein can modulate the activity of PKM2 and plays an important role in glycolytic metabolism.

USP25- ubiquitin specific peptidase 25; in conjunction with the 26S proteasome USP25 is involved in the ubiquitin-dependent proteolytic pathway.

PIAS4- protein inhibitor of activated STAT, 4; it functions as an E3-type SUMO ligase, which stabilize the interaction between the substrate and UBE2I. Hence it act as a SUMO-tethering factor. In various cellular pathways, it plays an important role as a transcriptional coregulation, including the the p53 pathway, STAT pathway, steroid hormone signalling pathway and the Wnt pathway. PIAS4 are involved in silencing of genes and also promotes PARK7 sumoylation.

UBE2I- ubiquitin-conjugating enzyme E2I: From the UBLE1A-UBLE1B E1 complex SUMO proteins are accepted by UBE2I. Hence by assistance of an E3 ligase such as RANBP2 or CBX4, UBE2I catalyses attachment of these proteins covalently with other proteins. It is also essential for chromosome segregation and nuclear architecture.

NFKBIA-: the activity of dimeric NF-kappa-B/REL complexes is inhibited due to covering of nuclear localization signals of NFKBIA. It becomes stimulated and got phosphorylated under immune and proinflammatory responses and promotes ubiquitination and later degradation.

SAE1- SUMO1 activating enzyme subunit 1: This enzyme is dimeric in structure and acts as an E1 ligase for SUMO1, as well as other SUMO isoforms. It faciliates activation of SUMO proteins and on SAE2 there is conserved cysteine residue along with a formed thioester.

G6PC2- glucose-6-phosphatase, catalytic, 2: it is responsible for hydrolysation of glucose-6-phosphate into glucose inside endoplasmic reticulum. It may be helps in glycogenolysis and gluconeogenesis which are responsible for glucose production.

Shortest Path Analysis:

a) Between SUMO proteins

SUMO1-SUMO2
SUMO1-SUMO3
SUMO1-SUMO4
SUMO2-SUMO3
SUMO2-SUMO4
SUMO3-VIM-SUMO4
SUMO3-SUMO2-SUMO4
SUMO3-SUMO1-SUMO4
SUMO3-PA2G4-SUMO4
SUMO3-UBC-SUMO4
SUMO3-JUN-SUMO4
SUMO3-FOS-SUMO4

b) Between SUMO1 and ERK1/MAPK3

-BLVRA-
-HIF1A-
-ESR1-
-HSF4-
-HSF1-
-GTF21-
-UBTF-
-KRT8-
-SP1-
-UBC-
-HIST1H4A-
-TP53-
-PTPRB-
-PPARA-
-STMN2-
-MAP2K1-
-TOP2B-
-BAZ1B-
-ELK1-
-HIS1H2AB-
-RXRA-
-ATP5B-
-SREBF2-
-SREBF1-
-CASP8-
-CEBPB-
-JUN-
-HDAC4-
-GATA1-
-NUP153-
-PFKM-

c) Between SUMO2 and HSP 27

-MFAP1-
-UBE21-
-SUMO1-
-SRRM2-
-YWHAZ-
-RIF1-
-YWHAQ-
-UBC-
- SAP18-
-USP1-
-CUL3-
-EIF4G1-
-SNCA-
-CYCS-
-HSF1-
-MYC-
-TP53-
-UCHL5-
-EFTUD2-
-BTBD12-
-DAXX-

d) Between SUMO2 and ERK1/MAPK3

-BAZ1B-
-NUP153-
-HDAC4-
-BLVRA-
-PFKM-
-TOP2B-
-HIF1A-
-GTF2A-
-HIST1H4A-
-SREBF2-
-SREBF1-
-ETS1-
-VBR5-
-RPS6KA4-
-TP53-
-CEBPB-
-UBC-
-SNCA-
-JUND-
-FASN-
-HSF1-
-STAT3-
-UBTF-
-KRT8-
-MAK14-
-JUN-
-MYC-

4. _Multiple sequence alignments of proteins under study:

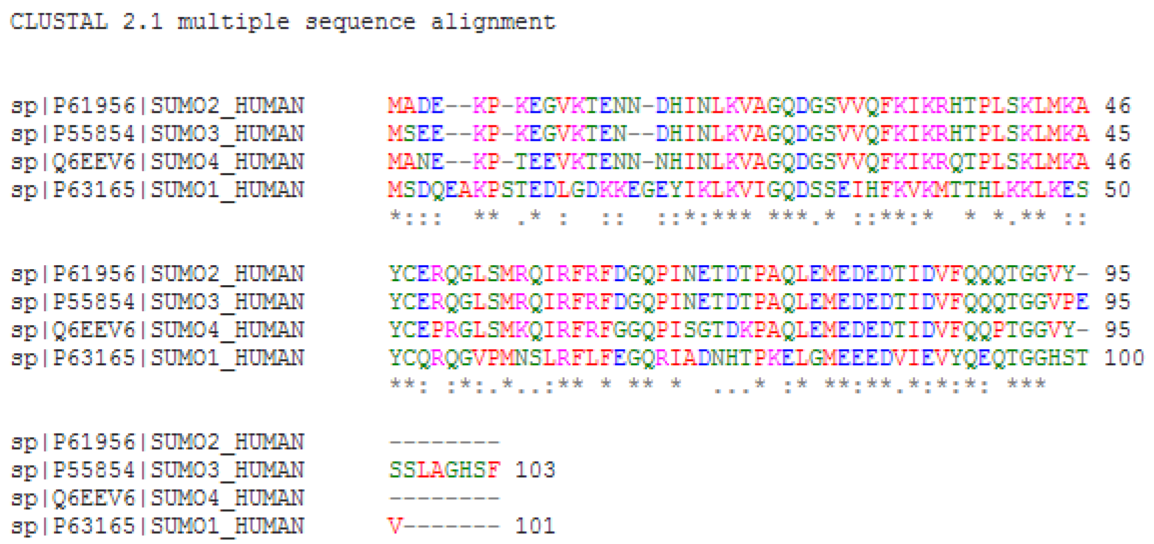
a) Between SUMO proteins:

Multiple sequence alignment was done with all four SUMO proteins with the help of Clustal Omega program. Results so obtained as output file is shown below:

Output file:

```
CLUSTAL 2.1 Multiple Sequence Alignments


Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: sp|P63165|SUMO1_HUMAN      101 AA
Sequence 2: sp|P61956|SUMO2_HUMAN       95 AA
Sequence 3: sp|P55854|SUMO3_HUMAN      103 AA
Sequence 4: sp|Q6EEV6|SUMO4_HUMAN       95 AA
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score: 43.16
Sequences (1:3) Aligned. Score: 43.56
Sequences (1:4) Aligned. Score: 38.95
Sequences (2:3) Aligned. Score: 95.79
Sequences (2:4) Aligned. Score: 86.32
Sequences (3:4) Aligned. Score: 83.16
Guide tree file created.

There are 3 groups
Start of Multiple Alignments

Aligning...
Group 1: Sequences:   2      Score:2002
Group 2: Sequences:   3      Score:1900
Group 3: Sequences:   4      Score:1517
Alignment Score 2236

CLUSTAL-Alignment file created.
```

Alignments :



```
CLUSTAL 2.1 multiple sequence alignment


sp|P61956|SUMO2_HUMAN      MADE--KP-KEGVKTENN-DHINLKVAGQDGSVVQFKIKRHTPLSKLMKA 46
sp|P55854|SUMO3_HUMAN      MSEE--KP-KEGVKTEN--DHINLKVAGQDGSVVQFKIKRHTPLSKLMKA 45
sp|Q6EEV6|SUMO4_HUMAN      MANE--KP-TEEVKTENN-NHINLKVAGQDGSVVQFKIKRQTPLSKLMKA 46
sp|P63165|SUMO1_HUMAN      MSDQEAKPSTEDLGDKKEGEYIKLKVIGQDSSEIHFKVKMTTHLKKLKES 50
                           *:::  ** .* :  ::  ::*:*** ***.* ::**:*  * *.** ::

sp|P61956|SUMO2_HUMAN      YCERQGLSMRQIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGGVY- 95
sp|P55854|SUMO3_HUMAN      YCERQGLSMRQIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGGVPE 95
sp|Q6EEV6|SUMO4_HUMAN      YCEPRGLSMKQIRFRFGGQPISGTDKPAQLEMEDEDTIDVFQQPTGGVY- 95
sp|P63165|SUMO1_HUMAN      YCQRQGVPMNSLRFLFEGQRIADNHTPKELGMEEEDVIEVYQEQTGGHST 100
                           **: :*:.*..:** * ** *  ...* :* **:**.*:*:*: ***

sp|P61956|SUMO2_HUMAN      --------
sp|P55854|SUMO3_HUMAN      SSLAGHSF 103
sp|Q6EEV6|SUMO4_HUMAN      --------
sp|P63165|SUMO1_HUMAN      V------- 101
```

Consensus Symbols:
"*" the residues or nucleotides in a particular column which are identical in all sequences.
":" conserved substitutions, according to the COLOUR table.
"." semi-conserved substitutions, i.e., amino acids having similar shape
Amino acid is replaced by another which have similar properties are conserved substitutions.

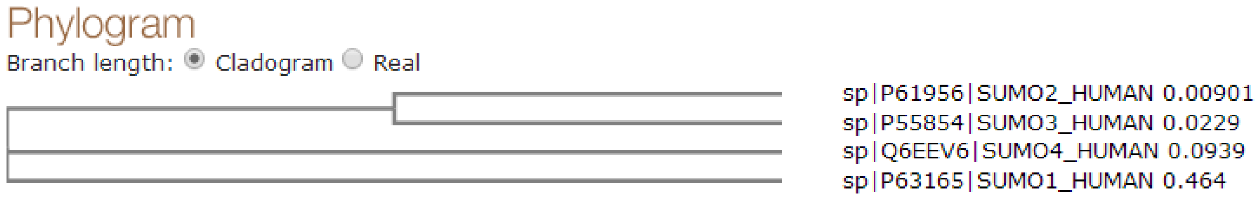
Score Table:

Seq A	Name	Length	Seq B	Name	Length	Score
1	sp|P63165|SUMO1_HUMAN	101	2	sp|P61956|SUMO2_HUMAN	95	43.16
1	sp|P63165|SUMO1_HUMAN	101	3	sp|P55854|SUMO3_HUMAN	103	43.56
1	sp|P63165|SUMO1_HUMAN	101	4	sp|Q6EEV6|SUMO4_HUMAN	95	38.95
2	sp|P61956|SUMO2_HUMAN	95	3	sp|P55854|SUMO3_HUMAN	103	**95.79**
2	sp|P61956|SUMO2_HUMAN	95	4	sp|Q6EEV6|SUMO4_HUMAN	95	86.32
3	sp|P55854|SUMO3_HUMAN	103	4	sp|Q6EEV6|SUMO4_HUMAN	95	83.16

Table 4: Score values in multiple sequence alignments

```
# Percent Identity Matrix - created by Clustal2.1
#
#
     1: sp|P61956|SUMO2_HUMAN  100.00   96.81   86.32   48.42
     2: sp|P55854|SUMO3_HUMAN  96.81  100.00   84.04   47.92
     3: sp|Q6EEV6|SUMO4_HUMAN  86.3284.04  100.00   44.21
     4: sp|P63165|SUMO1_HUMAN   48.42   47.92  44.21  100.00
```

On the basis of generated alignment scores and Percentage identity matrix above by Clustal Omega, a Phylogram is created which is showing distance relationship among SUMO proteins.



Phylogram

Branch length: ● Cladogram ○ Real

sp|P61956|SUMO2_HUMAN 0.00901
sp|P55854|SUMO3_HUMAN 0.0229
sp|Q6EEV6|SUMO4_HUMAN 0.0939
sp|P63165|SUMO1_HUMAN 0.464

<u>Analysis:</u>

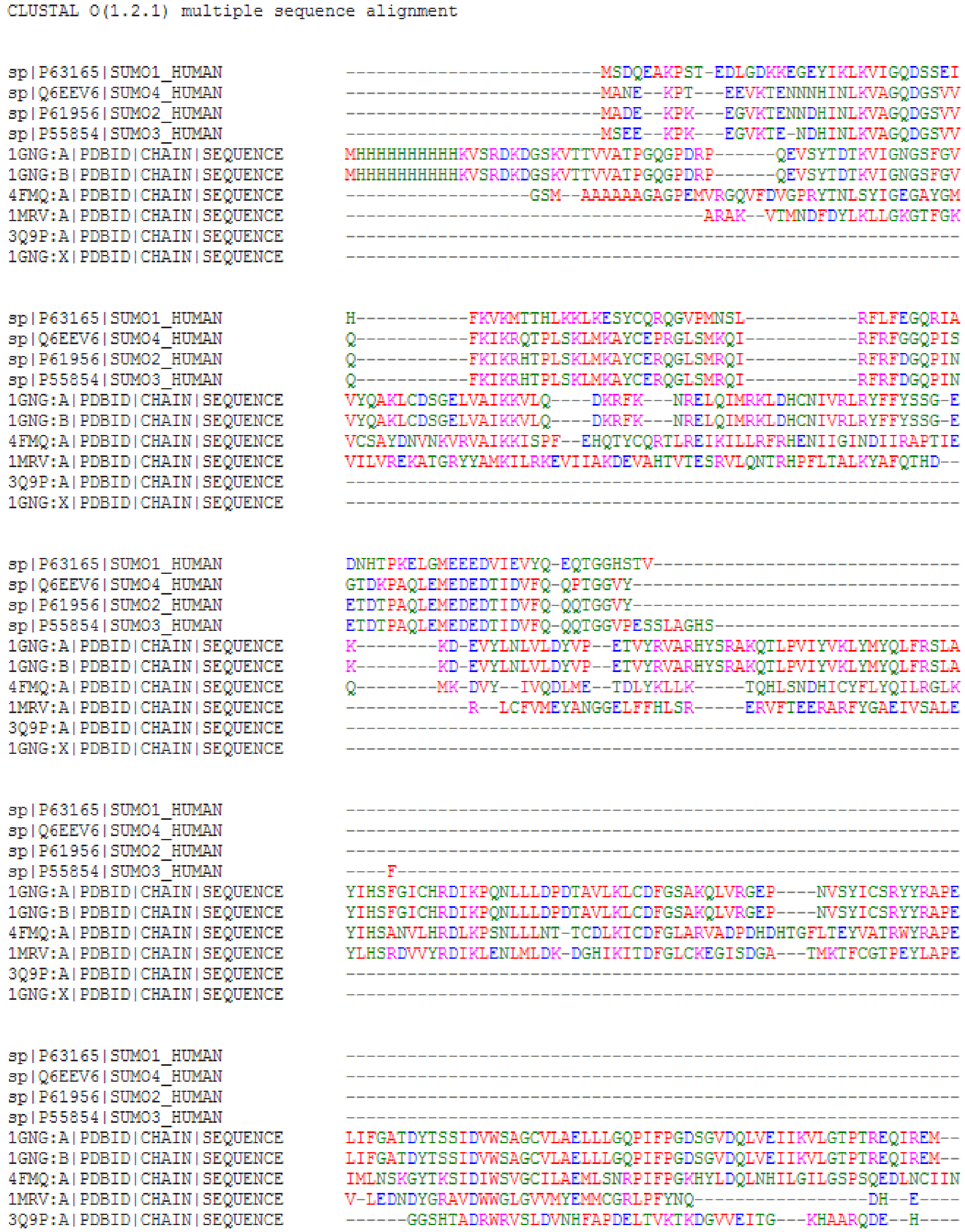
With the help of multiple alignments we can see which regions in proteins are conserved and which are not. Adding colours to amino acids helps us in identifying conserved regions easily.

Scores are obtained by pairwise alignments in MSA between different amino acid chains. Higher the score value in pairwise alignments higher will be the percentage identity between two aligning sequences. Hence from the score table we can say that SUMO 2 and SUMO 3 are very much similar to each other because of higher score value of about 95.79. This is crosschecked by looking into percentage identity matrix. Hence we can say that SUMO 2 and SUMO 3 are similar in structures and have same functions. They shows similar physical and chemical properties unlike SUMO 1 and SUMO 4 proteins having low score values.

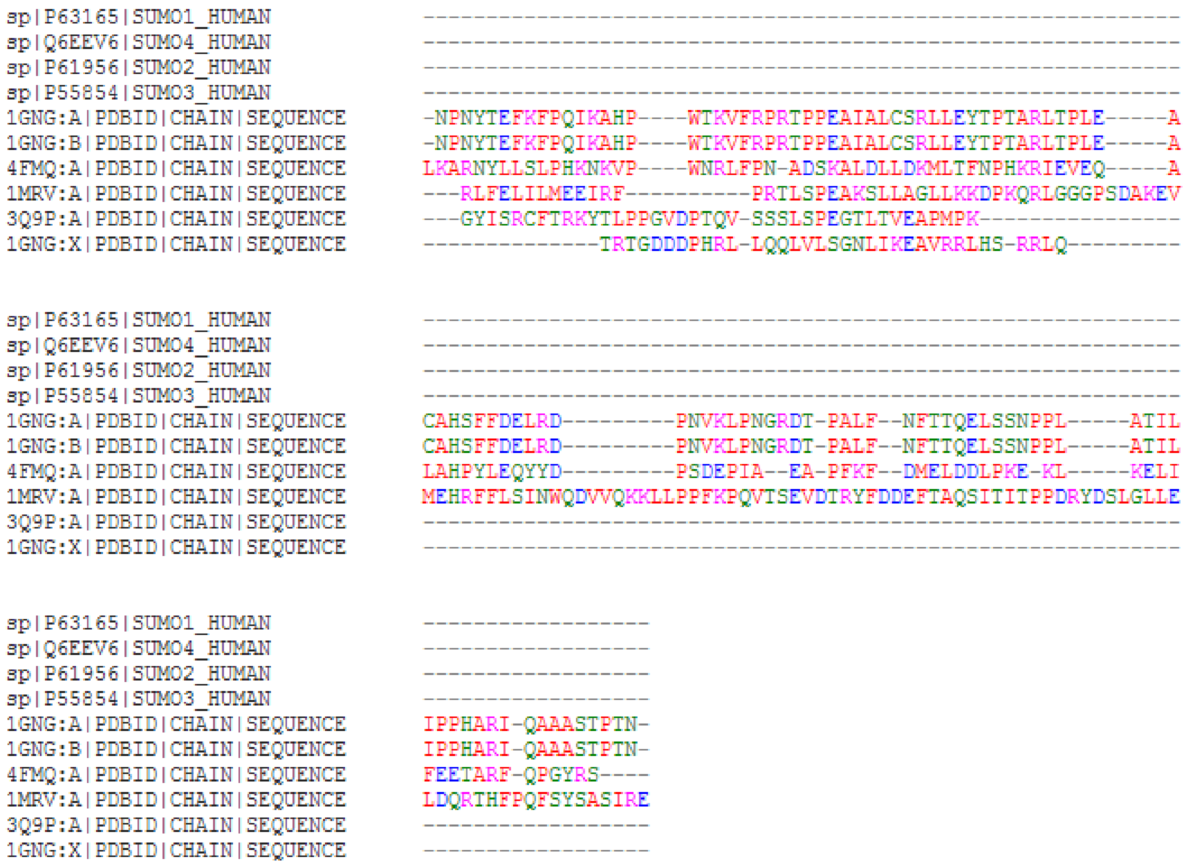
With the help of obtained data, a phylogram is drawn by using neighbour joining method which was shown in figure above. On the phylogram tree SUMO 2 and SUMO 3 are put on same branches because of their similar scores while SUMO 1 and SUMO 4 were put on other branches as per their scores.

b) Between SUMO proteins, GSK3, ERK2, AKT2 and HSP27:

Multiple sequence alignment was done with all four SUMO proteins and with GSK3, ERK2, AKT2 and HSP27 with the help of Clustal Omega program. Results so obtained as output file is shown below:



```
CLUSTAL O(1.2.1) multiple sequence alignment


sp|P63165|SUMO1_HUMAN         -------------------------MSDQEAKPST-EDLGDKKEGEYIKLKVIGQDSSEI
sp|Q6EEV6|SUMO4_HUMAN         -------------------------MANE--KPT---EEVKTENNNHINLKVAGQDGSVV
sp|P61956|SUMO2_HUMAN         -------------------------MADE--KPK---EGVKTENNDHINLKVAGQDGSVV
sp|P55854|SUMO3_HUMAN         -------------------------MSEE--KPK---EGVKTE-NDHINLKVAGQDGSVV
1GNG:A|PDBID|CHAIN|SEQUENCE   MHHHHHHHHHHKVSRDKDGSKVTTVVATPGQGPDRP------QEVSYTDTKVIGNGSFGV
1GNG:B|PDBID|CHAIN|SEQUENCE   MHHHHHHHHHHKVSRDKDGSKVTTVVATPGQGPDRP------QEVSYTDTKVIGNGSFGV
4FMQ:A|PDBID|CHAIN|SEQUENCE   ------------------GSM--AAAAAAGAGPEMVRGQVFDVGPRYTNLSYIGEGAYGM
1MRV:A|PDBID|CHAIN|SEQUENCE   -----------------------------------ARAK--VTMNDFDYLKLLGKGTFGK
3Q9P:A|PDBID|CHAIN|SEQUENCE   ------------------------------------------------------------
1GNG:X|PDBID|CHAIN|SEQUENCE   ------------------------------------------------------------

sp|P63165|SUMO1_HUMAN         H-----------FKVKMTTHLKKLKESYCQRQGVPMNSL-----------RFLFEGQRIA
sp|Q6EEV6|SUMO4_HUMAN         Q-----------FKIKRQTPLSKLMKAYCEPRGLSMKQI-----------RFRFGGQPIS
sp|P61956|SUMO2_HUMAN         Q-----------FKIKRHTPLSKLMKAYCERQGLSMRQI-----------RFRFDGQPIN
sp|P55854|SUMO3_HUMAN         Q-----------FKIKRHTPLSKLMKAYCERQGLSMRQI-----------RFRFDGQPIN
1GNG:A|PDBID|CHAIN|SEQUENCE   VYQAKLCDSGELVAIKKVLQ----DKRFK---NRELQIMRKLDHCNIVRLRYFFYSSG-E
1GNG:B|PDBID|CHAIN|SEQUENCE   VYQAKLCDSGELVAIKKVLQ----DKRFK---NRELQIMRKLDHCNIVRLRYFFYSSG-E
4FMQ:A|PDBID|CHAIN|SEQUENCE   VCSAYDNVNKVRVAIKKISPF--EHQTYCQRTLREIKILLRFRHENIIGINDIIRAPTIE
1MRV:A|PDBID|CHAIN|SEQUENCE   VILVREKATGRYYAMKILRKEVIIAKDEVAHTVTESRVLQNTRHPFLTALKYAFQTHD--
3Q9P:A|PDBID|CHAIN|SEQUENCE   ------------------------------------------------------------
1GNG:X|PDBID|CHAIN|SEQUENCE   ------------------------------------------------------------

sp|P63165|SUMO1_HUMAN         DNHTPKELGMEEEDVIEVYQ-EQTGGHSTV------------------------------
sp|Q6EEV6|SUMO4_HUMAN         GTDKPAQLEMEDEDTIDVFQ-QPTGGVY--------------------------------
sp|P61956|SUMO2_HUMAN         ETDTPAQLEMEDEDTIDVFQ-QQTGGVY--------------------------------
sp|P55854|SUMO3_HUMAN         ETDTPAQLEMEDEDTIDVFQ-QQTGGVPESSLAGHS------------------------
1GNG:A|PDBID|CHAIN|SEQUENCE   K--------KD-EVYLNLVLDYVP--ETVYRVARHYSRAKQTLPVIYVKLYMYQLFRSLA
1GNG:B|PDBID|CHAIN|SEQUENCE   K--------KD-EVYLNLVLDYVP--ETVYRVARHYSRAKQTLPVIYVKLYMYQLFRSLA
4FMQ:A|PDBID|CHAIN|SEQUENCE   Q--------MK-DVY--IVQDLME--TDLYKLLK-----TQHLSNDHICYFLYQILRGLK
1MRV:A|PDBID|CHAIN|SEQUENCE   ------------R--LCFVMEYANGGELFFHLSR-----ERVFTEERARFYGAEIVSALE
3Q9P:A|PDBID|CHAIN|SEQUENCE   ------------------------------------------------------------
1GNG:X|PDBID|CHAIN|SEQUENCE   ------------------------------------------------------------

sp|P63165|SUMO1_HUMAN         ------------------------------------------------------------
sp|Q6EEV6|SUMO4_HUMAN         ------------------------------------------------------------
sp|P61956|SUMO2_HUMAN         ------------------------------------------------------------
sp|P55854|SUMO3_HUMAN         ----F-------------------------------------------------------
1GNG:A|PDBID|CHAIN|SEQUENCE   YIHSFGICHRDIKPQNLLLDPDTAVLKLCDFGSAKQLVRGEP----NVSYICSRYYRAPE
1GNG:B|PDBID|CHAIN|SEQUENCE   YIHSFGICHRDIKPQNLLLDPDTAVLKLCDFGSAKQLVRGEP----NVSYICSRYYRAPE
4FMQ:A|PDBID|CHAIN|SEQUENCE   YIHSANVLHRDLKPSNLLLNT-TCDLKICDFGLARVADPDHDHTGFLTEYVATRWYRAPE
1MRV:A|PDBID|CHAIN|SEQUENCE   YLHSRDVVYRDIKLENLMLDK-DGHIKITDFGLCKEGISDGA---TMKTFCGTPEYLAPE
3Q9P:A|PDBID|CHAIN|SEQUENCE   ------------------------------------------------------------
1GNG:X|PDBID|CHAIN|SEQUENCE   ------------------------------------------------------------

sp|P63165|SUMO1_HUMAN         ------------------------------------------------------------
sp|Q6EEV6|SUMO4_HUMAN         ------------------------------------------------------------
sp|P61956|SUMO2_HUMAN         ------------------------------------------------------------
sp|P55854|SUMO3_HUMAN         ------------------------------------------------------------
1GNG:A|PDBID|CHAIN|SEQUENCE   LIFGATDYTSSIDVWSAGCVLAELLLGQPIFPGDSGVDQLVEIIKVLGTPTREQIREM--
1GNG:B|PDBID|CHAIN|SEQUENCE   LIFGATDYTSSIDVWSAGCVLAELLLGQPIFPGDSGVDQLVEIIKVLGTPTREQIREM--
4FMQ:A|PDBID|CHAIN|SEQUENCE   IMLNSKGYTKSIDIWSVGCILAEMLSNRPIFPGKHYLDQLNHILGILGSPSQEDLNCIIN
1MRV:A|PDBID|CHAIN|SEQUENCE   V-LEDNDYGRAVDWWGLGVVMYEMMCGRLPFYNQ-----------------DH--E----
3Q9P:A|PDBID|CHAIN|SEQUENCE   ------GGSHTADRWRVSLDVNHFAPDELTVKTKDGVVEITG---KHAARQDE--H----
```

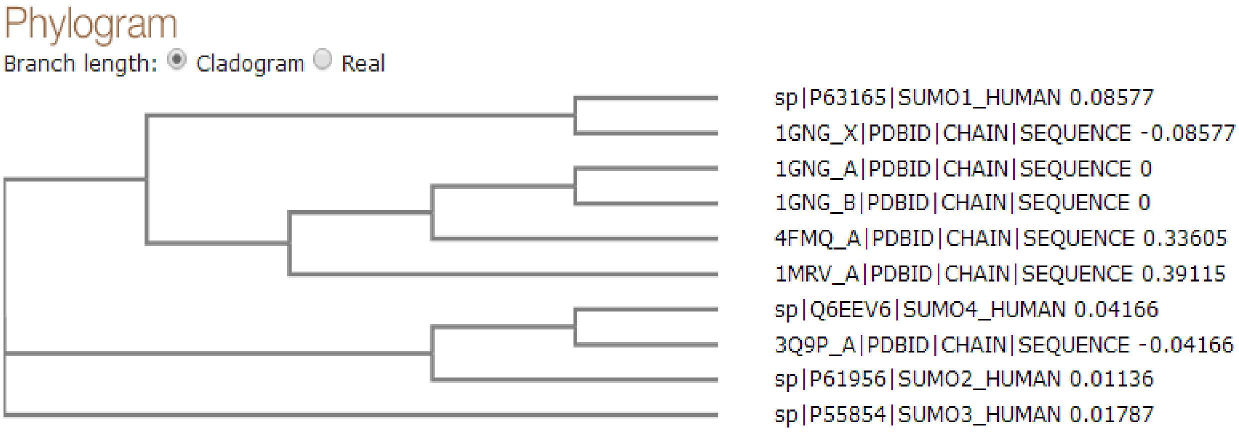


```
sp|P63165|SUMO1_HUMAN           ------------------------------------------------------------
sp|Q6EEV6|SUMO4_HUMAN           ------------------------------------------------------------
sp|P61956|SUMO2_HUMAN           ------------------------------------------------------------
sp|P55854|SUMO3_HUMAN           ------------------------------------------------------------
1GNG:A|PDBID|CHAIN|SEQUENCE     -NPNYTEFKFPQIKAHP----WTKVFRPRTPPEAIALCSRLLEYTPTARLTPLE-----A
1GNG:B|PDBID|CHAIN|SEQUENCE     -NPNYTEFKFPQIKAHP----WTKVFRPRTPPEAIALCSRLLEYTPTARLTPLE-----A
4FMQ:A|PDBID|CHAIN|SEQUENCE     LKARNYLLSLPHKNKVP----WNRLFPN-ADSKALDLLDKMLTFNPHKRIEVEQ-----A
1MRV:A|PDBID|CHAIN|SEQUENCE     ---RLFELILMEEIRF----------PRTLSPEAKSLLAGLLKKDPKQRLGGGPSDAKEV
3Q9P:A|PDBID|CHAIN|SEQUENCE     ---GYISRCFTRKYTLPPGVDPTQV-SSSLSPEGTLTVEAPMPK----------------
1GNG:X|PDBID|CHAIN|SEQUENCE     --------------TRTGDDDPHRL-LQQLVLSGNLIKEAVRRLHS-RRLQ---------

sp|P63165|SUMO1_HUMAN           ------------------------------------------------------------
sp|Q6EEV6|SUMO4_HUMAN           ------------------------------------------------------------
sp|P61956|SUMO2_HUMAN           ------------------------------------------------------------
sp|P55854|SUMO3_HUMAN           ------------------------------------------------------------
1GNG:A|PDBID|CHAIN|SEQUENCE     CAHSFFDELRD---------PNVKLPNGRDT-PALF--NFTTQELSSNPPL-----ATIL
1GNG:B|PDBID|CHAIN|SEQUENCE     CAHSFFDELRD---------PNVKLPNGRDT-PALF--NFTTQELSSNPPL-----ATIL
4FMQ:A|PDBID|CHAIN|SEQUENCE     LAHPYLEQYYD---------PSDEPIA--EA-PFKF--DMELDDLPKE-KL-----KELI
1MRV:A|PDBID|CHAIN|SEQUENCE     MEHRFFLSINWQDVVQKKLLPPFKPQVTSEVDTRYFDDEFTAQSITITPPDRYDSLGLLE
3Q9P:A|PDBID|CHAIN|SEQUENCE     ------------------------------------------------------------
1GNG:X|PDBID|CHAIN|SEQUENCE     ------------------------------------------------------------

sp|P63165|SUMO1_HUMAN           ------------------
sp|Q6EEV6|SUMO4_HUMAN           ------------------
sp|P61956|SUMO2_HUMAN           ------------------
sp|P55854|SUMO3_HUMAN           ------------------
1GNG:A|PDBID|CHAIN|SEQUENCE     IPPHARI-QAAASTPTN-
1GNG:B|PDBID|CHAIN|SEQUENCE     IPPHARI-QAAASTPTN-
4FMQ:A|PDBID|CHAIN|SEQUENCE     FEETARF-QPGYRS----
1MRV:A|PDBID|CHAIN|SEQUENCE     LDQRTHFPQFSYSASIRE
3Q9P:A|PDBID|CHAIN|SEQUENCE     ------------------
1GNG:X|PDBID|CHAIN|SEQUENCE     ------------------
```

1GNG: Glycogen Synthetase Kinase 3, 4FMQ: ERK 2 complex:

1MRV: AKT 2 Kinase domain, 3Q9P: HSP 27/HSP B1

On the basis of generated alignment scores by Clustal Omega, a Phylogram is created which is showing distance relationship among SUMO proteins.

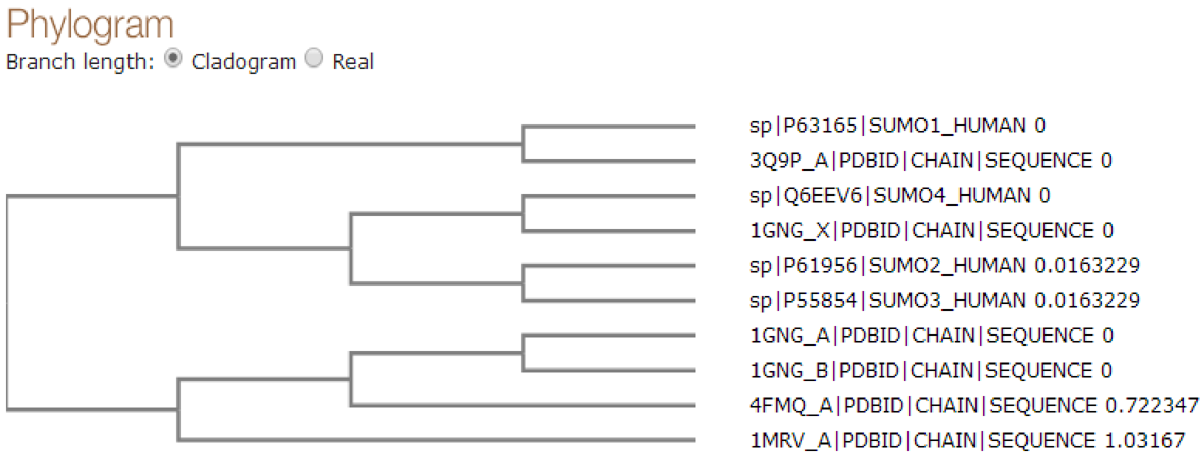


Phylogram

Branch length: ● Cladogram ○ Real

sp|P63165|SUMO1_HUMAN 0.08577
1GNG_X|PDBID|CHAIN|SEQUENCE -0.08577
1GNG_A|PDBID|CHAIN|SEQUENCE 0
1GNG_B|PDBID|CHAIN|SEQUENCE 0
4FMQ_A|PDBID|CHAIN|SEQUENCE 0.33605
1MRV_A|PDBID|CHAIN|SEQUENCE 0.39115
sp|Q6EEV6|SUMO4_HUMAN 0.04166
3Q9P_A|PDBID|CHAIN|SEQUENCE -0.04166
sp|P61956|SUMO2_HUMAN 0.01136
sp|P55854|SUMO3_HUMAN 0.01787

Phylogrm on the basis of Neighbour joining method



Phylogram

Branch length: ● Cladogram ○ Real

sp|P63165|SUMO1_HUMAN 0
3Q9P_A|PDBID|CHAIN|SEQUENCE 0
sp|Q6EEV6|SUMO4_HUMAN 0
1GNG_X|PDBID|CHAIN|SEQUENCE 0
sp|P61956|SUMO2_HUMAN 0.0163229
sp|P55854|SUMO3_HUMAN 0.0163229
1GNG_A|PDBID|CHAIN|SEQUENCE 0
1GNG_B|PDBID|CHAIN|SEQUENCE 0
4FMQ_A|PDBID|CHAIN|SEQUENCE 0.722347
1MRV_A|PDBID|CHAIN|SEQUENCE 1.03167

Phylogrm on the basis of UPGMA method

Analysis:

With the help of multiple alignments we can see which regions in proteins are conserved and which are not. Adding colours to amino acids helps us in identifying conserved regions easily.

With the help of obtained data, a phylogram is drawn by using neighbour joining method which was shown in figure above.

Proteins which show highest matching scores on multiple alignments data or on phylogram showing same distance values or closer in branches may have similar origins. It means these may adhere or affect the functions of other matching partner.

Therefore, SUMO 1 and SUMO 4 proteins may affect the functions of HSP 27 chaperone and GSK 3 complex while SUMO 2 and SUMO 3 proteins which are similar in properties may have little or no effect on these proteins. SUMO proteins may have no effect on the functions of ERK 2 and AKT 2 complexes because of farness on phylogram tree.

5. SUMO sites prediction:

Sumoylation is a unique process with specific characteristics. In a sumoylated protein, only a few of them out of many lysines (K), could be actual sumoylation sites. Sumoylation sites can determines by presence or absence of a consensus motif. Most of them follows consensus motif while some follows non-consensus region. Hence we can classify SUMO binding sites as: Type-1 Consensus binding sites and Type-2 Non-consensus binding sites.

Consensus motif: ΨKXE/D

Where, Ψ= large hydrophobic residue
K= Lysine residue
X= any residue
D/E= any acidic residue

SUMO sites prediction was performed on those proteins which are found to have direct or indirect links with glucose metabolism pathways. Results so obtained are listed in the table below.

Proteins	Uniprot ID	A.A.	Position	Peptide	Score	Cut-off	Type
SGLT-1	P13866	664					No SUMO bind. sites
SGLT-2	P31639	672	567	RHSKEER	2.897		Type-2 Nonconsensus
GLUT-2	P11168	524	255	EEVKAKQ	3.265	2.64	Type-2 Nonconsensus
GLUT-4	P14672	509	261 495	AELKDEK QEVKPST	1.953 2.868	0.13 2.64	Type-1 Consesus Type-2 Nonconsesus
GIP	P09681	153	108	ANRKEEE	2.765	2.64	Type-2 Nonconsesus
GLP-1	P06CA0	271					No SUMO bind. sites
Insulin	P01308	110					
Glucagon	P01275	180	180	TDRK***	3.279	2.64	Type-2 Nonconsesus
DPP-4	D27487	766	175 539	IYVKIEP KSSKKYPL	2.55 2.809	0.13 2.64	Type-1 Consesus Type-2 Nonconsesus
SLC30A8	Q8IWU4	369	241	IYFKPEY	0.332	0.13	Type-1 Consesus
TCF7L2	Q9NQB0	619	22 320 340 341 430 525	ISFKDEG PTVKQES QDSKKEE DSKKEEE KKRKRDK RDAKSQP	1.057 5.284 3.618 3.25 2.985 2.794	0.13 0.13 2.64 2.64 2.64 2.64	Type-1 Consesus Type-1 Consesus Type-2 Nonconsesus Type-2 Nonconsesus Type-2 Nonconsesus Type-2 Nonconsesus
ABCC8	Q09428	1581	205 1319	REVKPPE GLLKTEA	3.059 1.54	2.64 0.13	Type-2 Nonconsesus Type-1 Consesus
KCNJ11	Q14654	390					
CAPN10	Q9HCN6	672	426 671	HLWKVEK AVMKT**	0.289 3.044	0.13 2.64	Type-1 Consesus Type-2 Nonconsesus
IRS-1	P35568	1242					No SUMO bind. sites
Glycogen Synthatase	P54840	703					No SUMO bind. sites
RAD	Q9968	391					No SUMO bind. sites

HNF4A	P41235	474	126 470	AGMKKEA TITKQEV	1.882 3.603	0.13 2.64	Type-1 Consesus Type-2 Nonconsesus
GCK	P35557	465	12 15	EAAKKEK KKEKVEQ	1.137 3.397	0.13 2.64	Type-1 Consesus Type-2 Nonconsesus
HNF1A	P20823	631	155 226	TPMKTQK NPSKEER	2.647 3.397	2.64 2.64	Type-2 Nonconsesus Type-2 Nonconsesus
IPF1	P52945	283	202	MKWKKEE	0.972	0.13	Type-1 Consesus
HNF1B	P35680	557	42 161 258	FGVKLET TPMKTQK NPSKEER	1.372 2.647 3.485	0.13 2.64 2.64	Type-1 Consesus Type-2 Nonconsesus Type-2 Nonconsesus
NEUROD1	Q13562	356	286	FSFKHEP	1.716	0.13	Type-1 Consesus
CDKAL1	Q5VV42	579	124 334 404 413 508	NSIKKAQ MEMKREY PAAKMEQ AQVKKQR PLAKGEV	3.279 1.118 1.607 2.779 0.502	2.64 0.13 0.13 2.64 0.13	Type-2 Nonconsesus Type-1 Consesus Type-1 Consesus Type-2 Nonconsesus Type-1 Consesus
HHEX/IDE	Q03014	270	196 204 205	RRLKQEN QSNKKEE SNKKEEL	3.739 3.441 3.426	0.13 2.64 2.64	Type-1 Consesus Type-2 Nonconsesus Type-2 Nonconsesus
IGF2BP2	Q9Y6M1	599	299 497 505 509 583 599	NLKKIEH GKLKEEN FNPKEEV EEVKLEA QQVKQQE QRSK***	2.721 2.152 0.422 1.991 3.559 3.221	2.64 0.13 0.13 0.13 2.64 2.64	Type-2 Nonconsesus Type-1 Consesus Type-1 Consesus Type-1 Consesus Type-2 Nonconsesus Type-2 Nonconsesus
CDKN2A	P42771	156					No SUMO bind. sites
PPARC1A	Q07869	468	185	AKLKAEI	1.905	0.13	Type-1 Consesus
FOXC2	Q99958	501	214 227	VVIKSEA VITKVET	3.242 2.662	0.13 2.64	Type-1 Consesus Type-2 Nonconsesus
New diabetes genes identified by IGIB scientists							
GRB14	Q14449	540	193	NYAKYEF	0.232	0.13	Type-1 Consesus
ST6GAL1	P15907	406	88	AKAKPEA	1.161	0.13	Type-1 Consesus
VPS26A	Q75436	327	30 165 213 232 242	AEMKTED NSIKMEV QLIKKEI TIAKYEI APVKGES	2.128 2.872 2.057 0.36 1.891	0.13 0.13 0.13 0.13 0.13	Type-1 Consesus Type-1 Consesus Type-1 Consesus Type-1 Consesus Type-1 Consesus
HMG20A	Q9NP66	347	241	QLRKSNM	2.691	2.64	Type-2 Nonconsesus
AP3S2	P59780	193					No SUMO bind. sites
Proteins under study in the project							
HSP27	P04792	205	198 205	EAAKSDE TAAK***	3.191 3.794	2.64 2.64	Type-2 Nonconsesus Type-2 Nonconsesus
HSP70	P08107	641	325 512	RDAKLDK RLSKEEI	3.044 3.353	2.64 2.64	Type-2 Nonconsesus Type-2 Nonconsesus
AKT1	P31746	480	64 111 112 182 189 276 385	QLMKTER DGLKKQE GLKKQEE KILKKEV IVAKDEV RDLKLEN GLLKKDP	1.886 2.735 3.147 1.502 0.659 1.915 2.691	0.13 2.64 2.64 0.13 0.13 0.13 2.64	Type-1 Consesus Type-2 Nonconsesus Type-2 Nonconsesus Type-1 Consesus Type-1 Consesus Type-1 Consesus Type-2 Nonconsesus
AKT2	P31751	481	64 191 277	QLMKTER IIAKDEV RDIKLEN	1.886 0.55 2.735	0.13 0.13 0.13	Type-1 Consesus Type-1 Consesus Type-1 Consesus

			386	GLLKKDP	2.691	2.64	Type-2 Nonconsesus
ERK1	P27361	379	134	KLLKSQQ	2.985	2.64	Type-2 Nonconsesus
ERK2	P28482						No SUMO bind. sites

Table.5: SUMO binding sites prediction on key proteins involves directly or indirectly in glucose metabolism

Analysis:

SUMO binding sites are predicted by the tool based on presence or absence of consensus motif in the proteins. General scoring is done with cut-off as 0.13 and 0.27 for consensus and non-consensus binding sites. More the score value more will be the chance of predicted binding site to be true.

Proteins which shows higher score value have higher chances to be interacted by SUMO proteins. List of these proteins are: Glucagon, DPP-4, TCF7L2, HNF4A, HNF4B, HNF1B, CDKAL1, IDE, IGF2BP2, FOXC2, VPS26A

Also, proteins HSP27, AKT1 and AKT2 show higher score value which indicates that these proteins may interact with SUMO proteins.

Proteins which shows lower score values but above than cut-off scores may also get sumoylated which can be cross validated by wet laboratory experiments.

6. Motif and Domain analysis:

Proteins	**SUMO 1** **(101 aa)**	**SUMO 2** **(95 aa)**	**SUMO 3** **(103 aa)**	**SUMO 4** **(95 aa)**
Predicted features:				
DOMAIN	20-97, Ubiquitin like	16-93, Ubiquitin like	15-92, Ubiquitin like	16-93, Ubiquitin like
CROSSLINK	97, Gly-Lys isopeptide	93, Gly-Lys isopeptide	92, Gly-Lys isopeptide	93, Gly-Lys isopeptide
Patterns:				
N-myristoylation site	28-33: GQdsSE 56-61: GVpmNS 68-73: GQriAD 96-101: GGhsTV	24-29: GQdgSV 64-69: GQpiNE	23-28: GQdgSV 63-68: GQpiNE 92-97: GVpeSS	24-29: GQdgSV 64-69: GQpiSG
CAMP_phospho_site	--	35-38: KRhT	34-37: KRhT	35-38: KRqT
PKC_phospho_site	61-63: S1R MOD_RES 61, Phosphoserine 76-78: TpK MOD_RES 76, Phosphothreonine	54-56: SmR MOD_RES 54, Phosphoserine	53-55: SmR MOD_RES 53, phosphoserine	54-56: SmK MOD_RES 54, Phosphoserine 70-72: TdK MOD_RES 70, Phosphothreonine
N-glycosylation site	--	68-71: NETD CARBOHYD 68, N-linked (GlcNAc.)	67-70: NETD CARBOHYD 67, N-linked (GlcNAc.)	--
CK2_phospho_site	2-5: SdqE MOD_RES 2, Phosphoserine 9-12: SteD MOD_RES 9, phosphoserine 76-79: TpkE MOD_RES 76, Phosphothreonine	--	12-15: TenD MOD_RES 12, Phosphothreonine	68-71: SgtD MOD_RES 68, Phosphoserine

Table.6: Predicted Features and Patterns of SUMO proteins

7. Physio-chemical parameter computation:

To, understand more about the interacted proteins, the different kinds of physiochemical parameters are calculated using ProtParam.

Protein	**SUMO1**	**SUMO2**	**SUMO3**	**SUMO4**	**HSP27**	**AKT2**	**ERK2**
No. of amino acids	101	95	103	95	205	481	360
Molecular weight	11556.9	10871.2	11637.0	10685.1	22782.5	55768.7	41389.7
Ex. Coefficient	4470	2980	1490	2980	40450	67185	45185
Theoretical pI	5.34	5.32	5.32	6.57	5.98	5.98	6.50
Instability index	44.43	28.88	44.37	34.59	62.82	35.09	39.71
Aliphatic index	65.54	63.58	62.43	63.58	68.54	77.01	95.94
GRAVY	-0.916	-0.893	-0.828	-0.788	-0.567	-4.073	-0.287
Half life (hrs)	30	30	30	30	30	30	30
Stability	unstable	Stable	unstable	stable	unstable	stable	Stable

Table.7: Physico-chemical parameters of proteins

Analysis:

Proteins which have high instability index are unstable. Hence SUMO1, SUMO3 and HSP27 proteins are unstable while SUMO2, SUMO4, AKT2 and ERK2 correspond to low instability index are stable proteins.

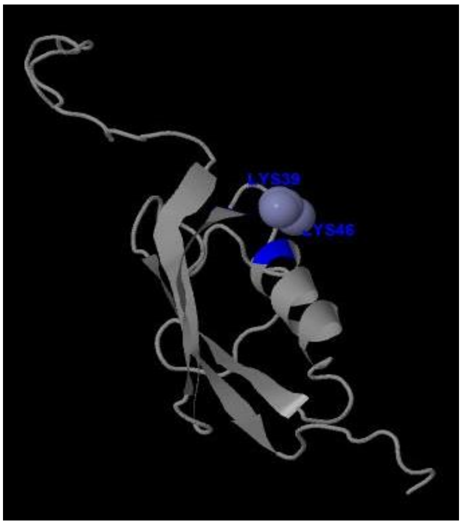
AKT2 have lowest GRAVY value among studied proteins. It means hydrophobicity is very less in this protein and can function efficiently in aqueous medium than others.

8. Binding sites prediction:

1. SUMO 1

Ligand Clusters Identified					
			MAMMOTH Scores		
Cluster	Ligands	Structures	Av	min	max
1	4	4	10.7	10.3	11.0
2	3	3	10.7	10.3	10.9
3	2	2	9.8	9.5	10.2
4	1	1	11.5	11.5	11.5
5	1	1	11.3	11.3	11.3
6	1	1	10.6	10.6	10.6
9	1	1	10.6	10.6	10.6
8	1	1	9.6	9.6	9.6
7	1	1	9.6	9.6	9.6

Binding Sites Prediction			
Residue	Amino acid	contact	Average distance
39	LYS	4	0.07
46	LYS	4	0.00

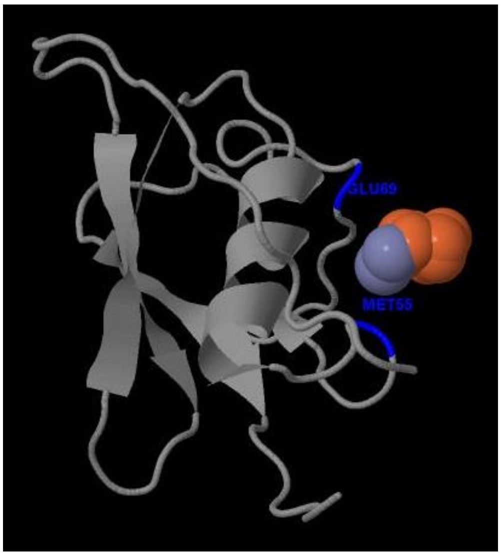


Heterogens present in Predicted Binding Site		
Heterogen	Count	source structures
ZN	4	2zcb_A, 2w9n_A, 2zcc_A, 3hm3_B

2. <u>SUMO 2</u>

Ligand Clusters Identified					
			MAMMOTH Scores		
Cluster	Ligands	Structures	Av	min	max
1	12	12	7.9	7.0	10.7
2	4	4	11.0	10.6	11.3
3	4	4	8.0	7.8	8.5
4	3	3	10.9	10.5	11.2
5	2	2	9.1	7.9	10.3
11	1	1	10.8	10.8	10.8
10	1	1	10.1	10.1	10.1
9	1	1	8.5	8.5	8.5
8	1	1	7.9	7.9	7.9
6	1	1	7.4	7.4	7.4
7	1	1	7.4	7.4	7.4

Binding Sites Prediction			
Residue	Amino acid	contact	Average distance
55	MET	8	0.34
69	GLU	12	0.00

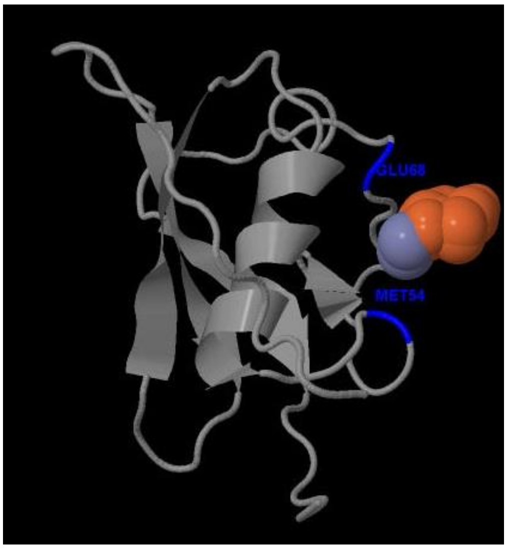


Heterogens present in Predicted Binding Site		
Heterogen	Count	source structures
FES	11	1nen_B, 1nek_B, 2wdv_F, 2wdr_B, 2wdq_F, 2acz_B, 2bs3_B, 2bs2_E, 1qlb_E, 1e7p_H, 2bs4_E
ZN	2	2w9n_A,3h7s_B

3. SUMO 3

Ligand Clusters Identified					
			MAMMOTH Scores		
Cluster	Ligands	Structures	Av	min	max
1	13	13	8.0	7.1	10.7
2	10	10	7.6	7.1	8.3
3	5	5	8.1	7.8	8.5
4	4	4	11.0	10.6	11.3
5	3	3	10.9	10.5	11.2
6	2	2	8.0	8.0	8.0
9	1	1	10.8	10.8	10.8
8	1	1	10.8	10.8	10.8
7	1	1	8.2	8.2	8.2

Binding Sites Prediction			
Residue	Amino acid	contact	Average distance
54	MET	8	0.24
68	GLU	13	0.00

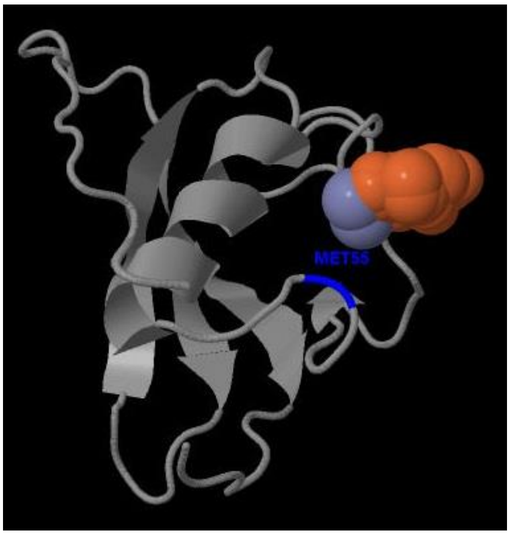


Heterogens present in Predicted Binding Site		
Heterogen	Count	source structures
FES	11	1nen_B, 1nek_B, 2wdv_F, 2wdr_B, 2wdq_F, 2acz_B, 2bs3_B, 2bs2_E, 1qlb_E, 1e7p_H, 2bs4_E
ZN	2	2w9n_A,3h7s_B

4. SUMO 4

Ligand Clusters Identified					
			MAMMOTH Scores		
Cluster	Ligands	Structures	Av	min	max
1	12	12	7.9	7.0	10.7
3	4	4	11.0	10.6	11.3
2	4	4	8.0	7.8	8.5
4	3	3	10.9	10.5	11.2
5	2	2	9.1	7.9	10.3
11	1	1	10.8	10.8	10.8
10	1	1	10.1	10.1	10.1
9	1	1	8.5	8.5	8.5
8	1	1	7.9	7.9	7.9
6	1	1	7.4	7.4	7.4
7	1	1	7.4	7.4	7.4

Binding Sites Prediction			
Residue	Amino acid	contact	Average distance
55	MET	8	0.34



Heterogens present in Predicted Binding Site		
Heterogen	Count	source structures
FES	10	1nen_B, 1nek_B, 2wdq_F, 2acz_B, 2wdr_B, 2bs4_E, 2bs3_B, 2bs2_E, 1qlb_E, 1e7p_H
ZN	2	2w9n_A,3h7s_B

Analysis:

The Cluster Table

The details of the clusters of ligands can be determined on the target model with the help of this table. There are 9 clusters found for Sumo1 and Sumo3 while 11 clusters found for Sumo 2 and Sumo4 proteins.

Predicted Binding Site Table

This table shows amino acids which are predicted to form part of the binding site. The residue number and amino acid code is shown for every residue in the table. The number of ligands that are in contact or close to the residue are also displayed in the table.

SUMO proteins have 2 binding sites which are positioned at: 39 Lys and 46 Lys in SUMO1 while 55 Met and 69 Glu in SUMO2, and 54 Met and 68 Glu in SUMO3 proteins.

SUMO 4 protein contains only 1 binding sites positioned at 55wt.

Heterogen Table

Heterogens present in the ligand cluster are listed in heterogen table. The structures from which each type of ligand originates along with numerosity of ligand in them are also presented in the table.

FeS and Zn type heterogens/ligands are found to show interactions with Sumo2, Sumo3 and Sumo4 while Sumo1 shows interaction with only FeS kind of heterogens.

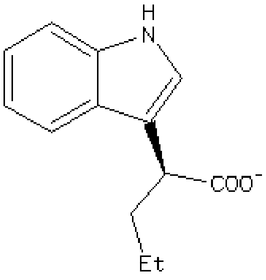
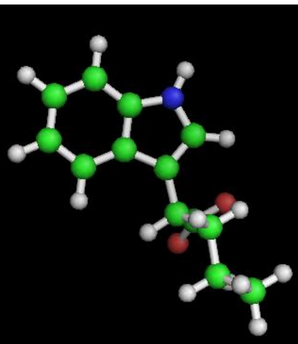
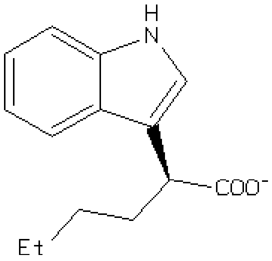
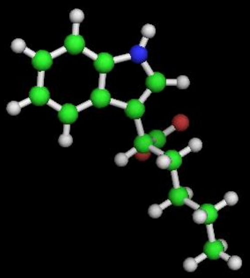
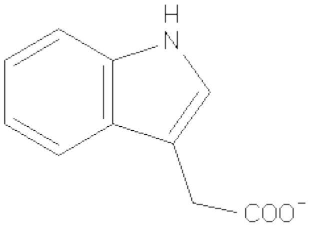
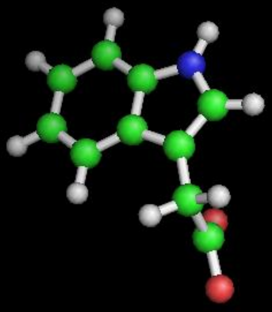
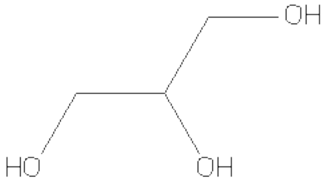
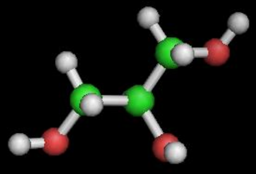
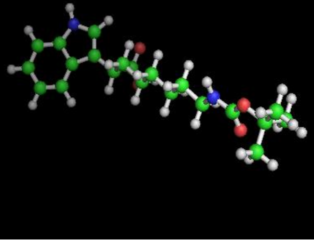
Figures are displaying the modelled structure of SUMO proteins with ligands in binding sites. The residues predicted to form part of the binding site are coloured blue.

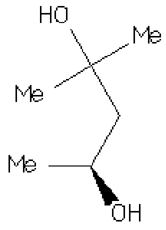
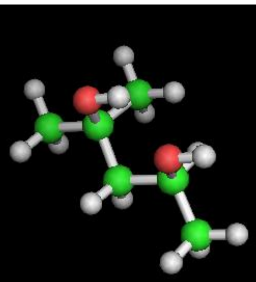
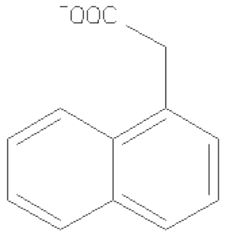
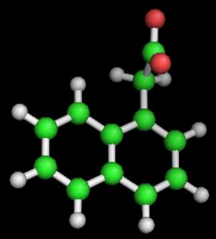
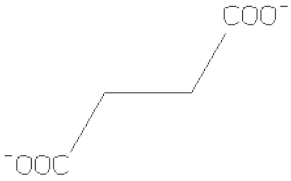
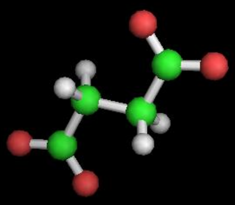
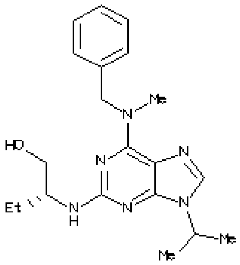
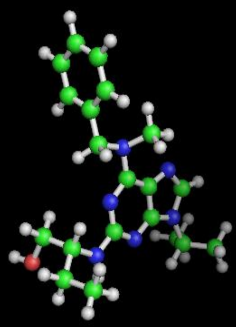
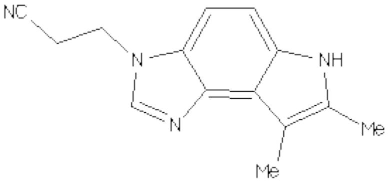
9. Virtual Screening of ligands from ZINC database:

S. No.	ZINC ID/ Drug Bank ID	Matthew's correlation coefficient (Probability Score)	Compound Name
1.	ZINC39119680 DB06981	0.977069	(2S)-2-(1H-indol-3-yl) pentanoic acid
2.	ZINC53683750 DB06980	0.976618	(2S)-2-(1H-indol-3-yl) hexanoic acid
3.	ZINC00083860 DB07950	0.972639	1H-indol-3-yl acetic acid
4.	ZINC00895048 DB04077	0.969267	Glycerol
5.	ZINC53683754 DB06982	0.965487	(2S)-8-[(tert-butoxycarbonyl)amino]-2-(1H-indol-3-yl)octanoic acid]
6.	ZINC00388075 DB03564	0.954248	(4R)-2-methylpentane-2,4-diol
7.	ZINC00391809 DB01750	0.950492	naphthalen-1-yl-acetic acid
8.	ZINC00895030 DB00139	0.943693	Succinic acid
9.	ZINC12504508 DB04776	0.923399	N6-methyl-(R)-roscovitne, R-2-[6-(benzyl-methyl-amino)-9-isopropyl-9H-purin-2-ylamino]-butan-1-ol
10.	ZINC01802814 --	0.921811	3-(7,8-dimethylpyrrolo[3,2-e]benzimidazol-3-yl) propanenitrile
11.	ZINC00643153 DB01026	0.921448	Ketoconazole
12.	ZINC19796080 DB00450	0.921429	Droperidol
13.	ZINC00001673 DB01042	0.921425	Melphalan
14.	ZINC00000797 DB07615	0.920838	2-{[(2E)-3-(3,4-dimethoxyphenyl)prop-2-enoyl]amino}benzoic acid
15.	ZINC03779067 DB04115	0.920741	Berberine
16.	ZINC01851184 --	0.920691	(6ar,12bS)-5,6,6a,7,8,12b-hexahydrobenzo [a]phenanthridin-6-ium-10,11-diol
17.	ZINC02021799 DB00409	0.920637	Remoxipride
18.	ZINC00215680 --	0.920426	8,8-dimethyl-5-morpholino-8,9-dihydro-6H-isothaizolo[5,4-b]pyrano[4,3-d]pyridine-1-amine
19.	ZINC00537805 DB01016	0.920424	Glyburide
20.	ZINC00057480 DB00867	0.920287	Ritodrine

Table.8: Top 20 virtually screened compounds against SUMO1 receptor

Figure.7: ZINC Accession no. and structures of top 10 virtually screened compounds:

S. No.	ZINC Acc. No./ Compound name	Structure	
1	ZINC39119680 **(2S)-2-(1H-indol-3-yl)pentanoic acid**		
2	ZINC53683750 **(2S)-2-(1H-indol-3-yl)hexanoic acid**		
3	ZINC00083860 **Indole-3-acetic acid**		
4	ZINC00895048 **Glycerol**		
5	ZINC53683754 **(2S)-8-[(tert-butoxycarbonyl)amino]-2-(1H-indol-3-yl)octanoic acid**		

6	ZINC00388075 **2-methylpentane-2,4-diol**		
7	ZINC00391809 **1-Naphthylacetic acid**		
8	ZINC00895030 **Succinic acid**		
9	ZINC12504508 **N6-methyl-(R)-Roscovitine, R-2-[6-(benzyl-methyle-amino)-9-isopropyl-9H-purin-2-yl amino]-butan-1-oL**		
10	ZINC01802814 **3-(7,8-dimethylimidazo[4,5-e]indol-3(6H)-yl)propanenitrile**		

10.Docking Results:

Docking tools PatchDock and SwissDock were used to perform docking operations between receptor protein and top 10 virtually screened ligands.

As all four SUMO proteins are found to have near about same motifs and domain, I have used only SUMO1 protein to performed docking with virtually screened ligands with the use of docking tools.

The docking results are shown in the table below:

S. No.	Ligands (ZINC Acc. No.)	Docking with Receptor SUMO1				
		PatchDock Results			SwissDock Results	
		Score	Area	ACE	Fullfitness	Estimated ΔG
1.	ZINC 03911680	3012	352.20	-143.35	-708.30	-7.33
2.	ZINC 53683750	3072	365.80	-184.77	-710.44	-7.58
3.	ZINC 00083860	2554	285.00	-114.13	-703.45	-7.33
4.	ZINC 00895048	1760	201.60	-50.64	-670.46	-6.03
5.	ZINC 00388075	2256	239.00	-89.72	-699.62	-6.65
6.	ZINC 01802814	3292	427.20	-146.26	-686.00	-6.50
7.	ZINC 53683754	4892	578.00	-199.27	-756.75	-8.33
8.	ZINC 00391809	2602	313.80	-97.62	-692.33	-7.51
9.	ZINC 00895030	1772	216.20	-54.13	-746.84	-8.00
10.	ZINC 12504508	4054	599.50	-284.20	-702.82	-6.85

Table.9: Docking results of top 10 virtually screened ligands with SUMO1 receptor protein.

Analysis of docking results:

The aim of using two docking tools was to perform comparative docking analysis. Results from multiple docking runs are summarized in the table.

ParDock Result: Docking poses generated by the ParDock can be directly loaded into PyMOL. Information containing the docking score is displayed in the Table, allowing direct analysis of configuration/score relationships. The docked ligands were ranked according to their scores and their corresponding binding poses may be exported.

- **Score:** Geometric shape complementarity score. The solutions are sorted according to this score.
- **Area:** Approximate interface area of the complex.
- **ACE:** Atomic contact energy according to Zhang et al.

SwissDock Result:

SwissDock performs ranking of docked ligands according to their Fulftness score and ΔG, free energy of binding.

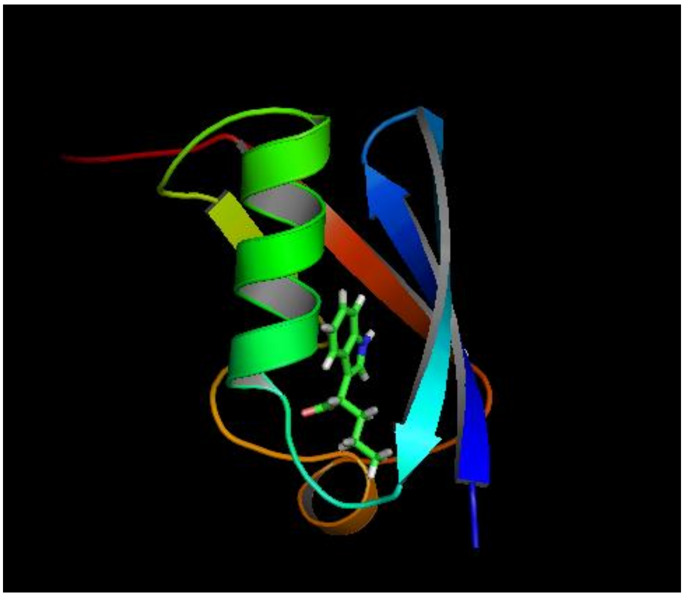
All docking poses in the population are assigned a fitness regarding a selection pressure or Simple Fitness based on their energy, and the fittest ones are automatically incorporated in the next generation.

Gibbs free energy, ΔG is the criteria used by SwissDock in predicting best docked ligands according to their scores. Lowest the free energy value of docked ligand, the more it will be preferred as interacting compound with the protein.

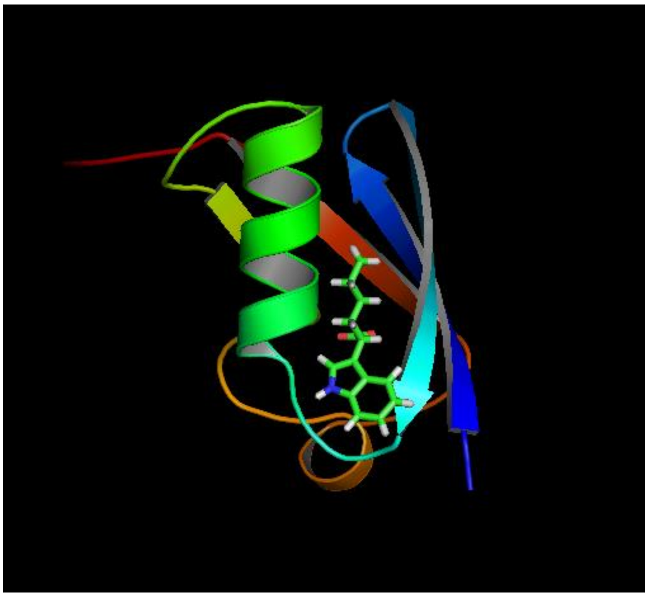
In PatchDock docking Ligands with ZINC ids: ZINC53683754, ZINC 12504508, ZINC01802814, ZINC53683750 were found to have higher scores while Ligands with ZINC ids: ZINC53683754, ZINC00895030, ZINC53683750 showed higher fitness scores and ΔG values. This shows that both Docking algorithms generate different results as per different criterias used by them for scoring which further question their credentiallity. Hence the best docked ligands can be predicted by analysing docked ligands by looking onto their interactions in the binding sites, their surface area of interaction and kinds of residues or atoms involved in the interactions. We can also predict best docking by performing extra calculations on the most promising complexes using atomistic models (eg.charmm, gromacs).

We found that complexes having ligands with ZINC ids: **ZINC53683754** [(2S)-8-{(tert-butoxycarbonyl)amino}-2-(1H-indole-3-oyl)octanoic acid] and **ZINC53683750** [(2S)-2-(1H-indole-3-yl)hexanoic acid] had highest scores. This result is supported by both PatchDock and SwissDock tools. Hence we can say that both these ligands are good binders and they can be used as agonist/antagonist against SUMO proteins.

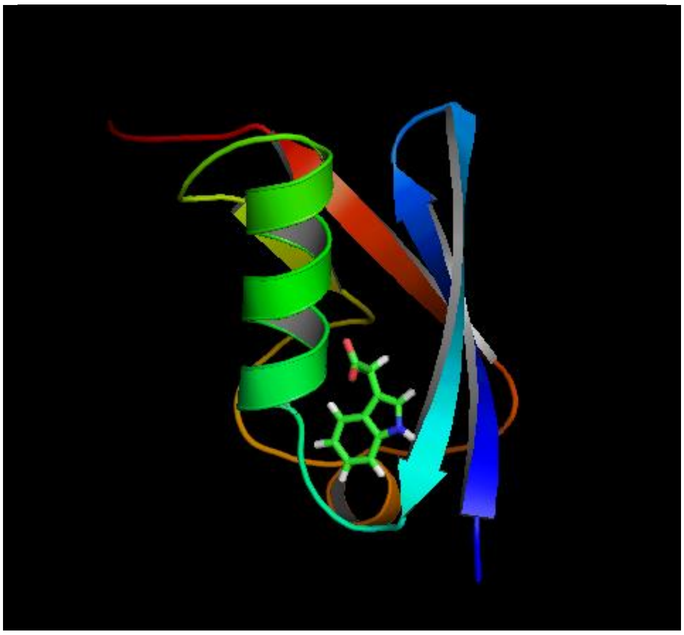
Figure.9: PyMol images of docked ligands within binding sites of SUMO1 receptor:



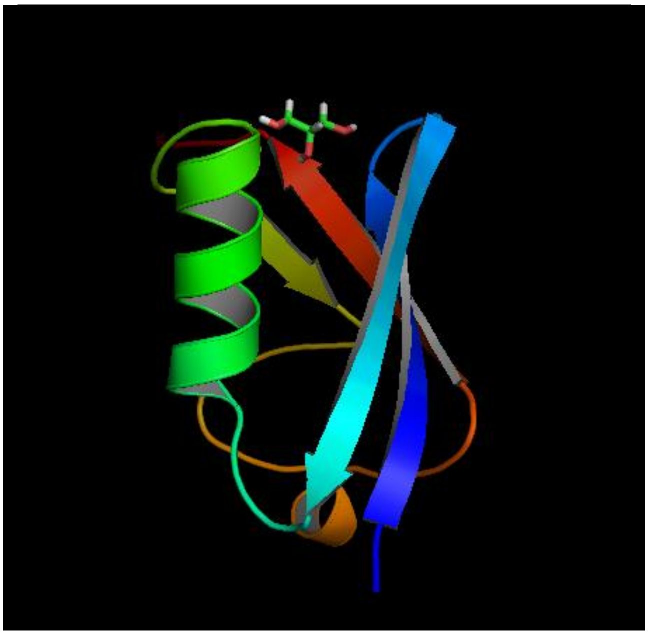
ZINC39119680
(2S)-2-(1H-indol-3-yl)pentanoic acid



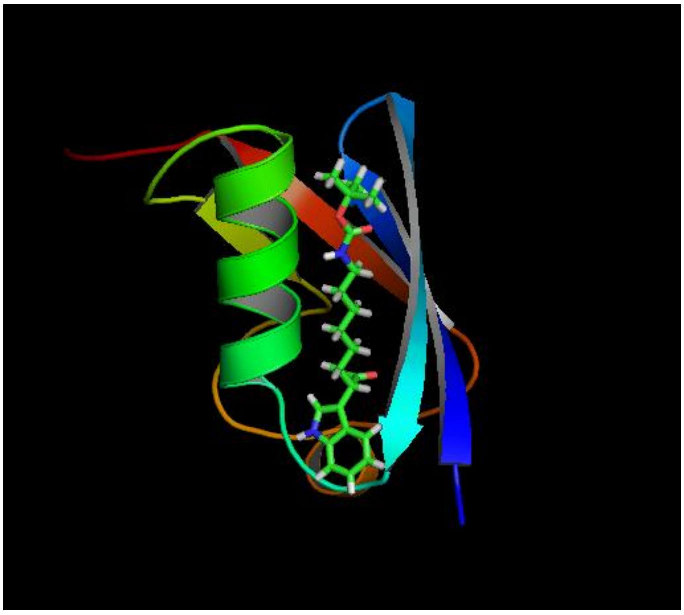
ZINC53683750
(2S)-2-(1H-indol-3-yl)hexanoic acid



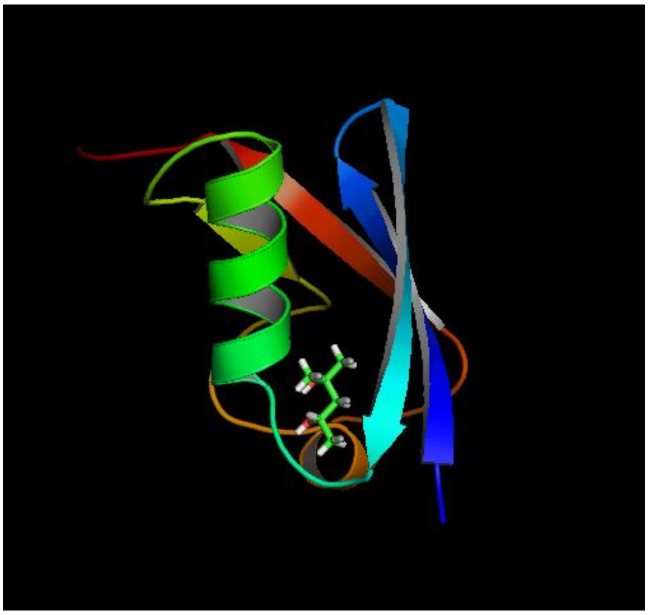
ZINC00083860
Indole-3-acetic acid



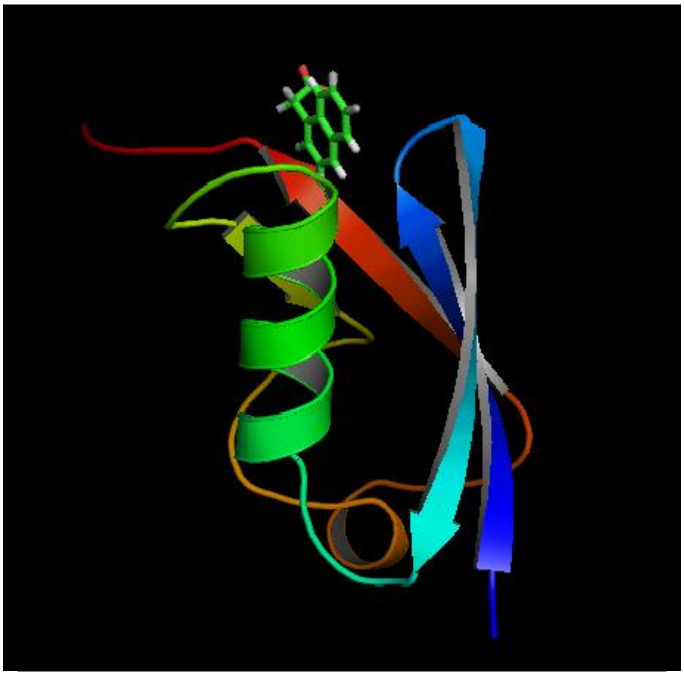
ZINC00895048
Glycerol



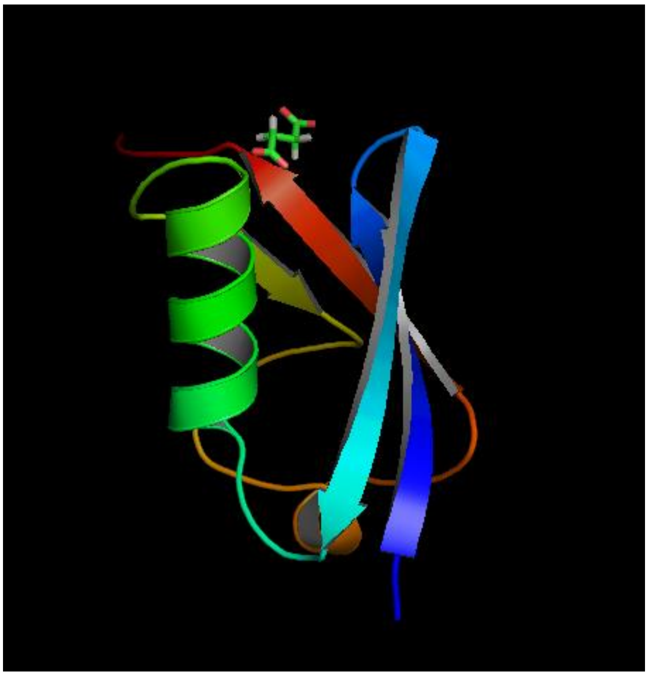
ZINC53683754
(2S)-8-[(tert-butoxycarbonyl)amino]-2-(1H-indol-3-yl)octanoic acid



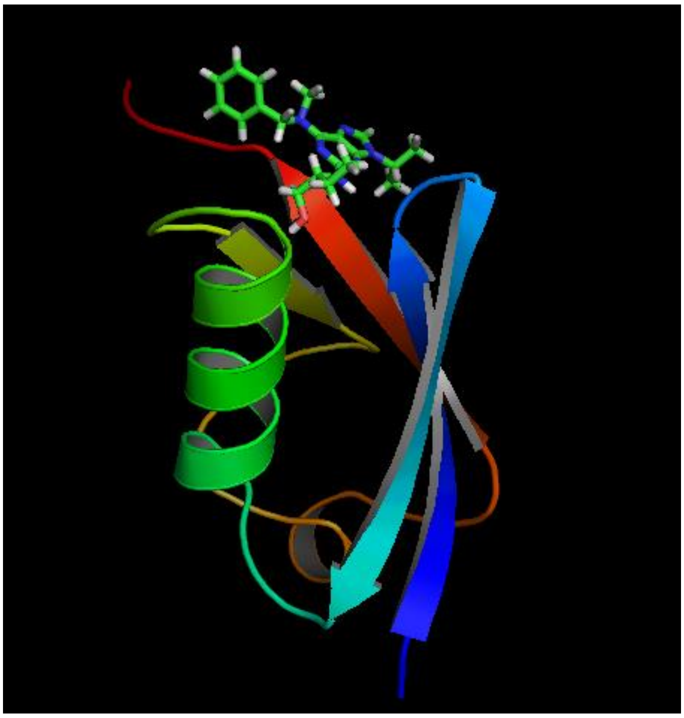
ZINC00388075
2-methylpentane-2,4-diol



ZINC00391809
1-Naphthylacetic acid

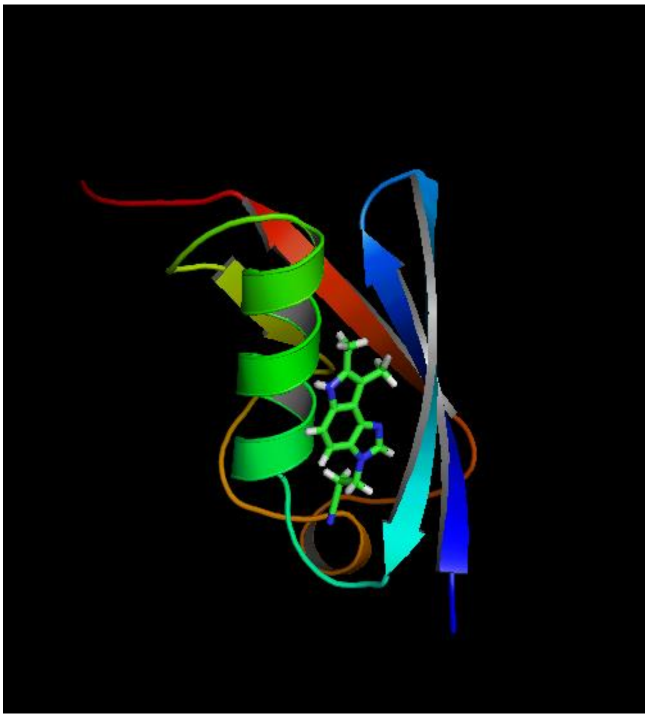


ZINC00895030
Succinic acid



ZINC12504508

N6-methyl-(R)-Roscovitine, R-2-[6-(benzyl-methyle-amino)-9-isopropyl-9H-purin-2-yl amino]-butan-1-oL



ZINC01802814

3-(7,8-dimethylimidazo[4,5-e]indol-3(6H)-yl)propanenitrile

CONCLUSION

It was found that SUMO proteins play important role in biological process and have important functions.

SUMO site prediction analysis was performed on a various proteins showing direct or indirect relationship with glucose metabolism pathway, it was found that proteins which got significantly higher scores may have SUMO binding sites e.g.Glucagon, DPP-4, TCF7L2, HNF4A, HNF4B, HNF1B, CDKAL1, IDE, IGF2BP2, FOXC2, VPS26A. Proteins HSP27, AKT1 and AKT2 also show higher score value which indicates that these proteins may interact with SUMO proteins. And, if these SUMO binding sites are cross validated by wet laboratory analysis then we can say that presence or absence of SUMO proteins in their working zones may alter their functions.

From protein-protein interaction analysis different interacted proteins are obtained which shows that altering of one protein may alter the functioning of other interacted protein.

In physio-chemical properties analysis of proteins, SUMO1, SUMO3 and HSP27 proteins are unstable while SUMO2, SUMO4, AKT2 and ERK2 correspond to low instability index are stable proteins.

AKT2 have lowest GRAVY value among studied proteins which make it less hydrophobic.

FeS and Zn type heterogens/ligands are found to show interactions with Sumo2, Sumo3 and Sumo4 while Sumo1 shows interaction with only FeS kind of heterogens.

SUMO proteins have 2 binding sites which are positioned at: 39 Lys and 46 Lys in SUMO1 while 55 Met and 69 Glu in SUMO2, and 54 Met and 68 Glu in SUMO3 proteins. SUMO 4 protein contains only 1 binding sites positioned at 55wt.

The uses of two or more docking tools are effective in efficient lead molecule prediction.This study suggest that **ZINC53683754** [(2S)-8-{(tert-butoxycarbonyl)amino}-2-(1H-indole-3-oyl)octanoic acid] and **ZINC53683750** [(2S)-2-(1H-indole-3-yl)hexanoic acid] had highest scores and can be used as a lead molecule. This result is supported by both PatchDock and SwissDock tools. Hence we can say that both these ligands are good binders and they can be used as agonist/antagonist against SUMO proteins for performing in vitro and in vivo study.

We conclude that drug discovery process can be speed up with the help of bioinformatics tools. These tools can also be helpful in cost cutting and may change the way of designing the drugs. Many naturally occurring chemical compounds found in herbs or medicinal plants can also be tested in silico for drug designing for finding new effective drug against diabetes, one of the top most killer disease.

DISCUSSION

For specific proteins SUMO makes covalent attachment with certain residues. This results in alteration of different functions of these proteins.

SUMO and Ubiquitinin competes for the same lysine residue in substrates. Hence proteosomal degradation which is characteristics of ubiquitination can be counteracted by sumoylation. In addition to this sumoylation is involved in subcellular localization, activation of transcription factors, DNA-binding and many other cellular functions.

Different docking algorithms generate different results as per different criterias used by them for scoring which further question their credentiallity. Hence the best docked ligands can be predicted by analysing docked ligands by looking onto their interactions in the binding sites, their surface area of interaction and kinds of residues or atoms involved in the interactions. We can also predict best docking by performing extra calculations on the most promising complexes using atomistic models (eg.charmm, gromacs).

Docking results of various chemical components from ZINC and DrugBank by PatchDock and Swissdock are listed in table 1 and 2 respectively. Use of two docking tools helps in finding efficient binders and also helps in performing comparative analysis. Results in Swissdock are found in terms of ΔG Kcal/mol and full fitness.

The best full fitness from Swissdock study was found for ZINC53683754 [(2S)-8-{(tert-butoxycarbonyl)amino}-2-(1H-indole-3-oyl)octanoic acid] andZINC53683750 [(2S)-2-(1H-indole-3-yl)hexanoic acid]

Recent reports indicate that regulation of sumo-conjugation contributes to the pathogenesis and development of cardiovascular complications.

REFRENCES

Ahner A, Gong X, Schmidt BZ, Peters KW, Rabeh WM, Thibodeau PH, Lukacs GL, Frizzell RA. *Small heat shock proteins target mutant CFTR for degradation via a SUMO-dependent pathway*.Department of Cell Biology and Physiology, University of Pittsburgh School of Medicine

Alkuraya FS, Saadi I, Lund JJ, Turbe-Doan A, Morton CC, Maas RL. *SUMO1 haploinsufficiency leads to cleft lip and palate*.Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, NRB-458, 77 Louis Pasteur, Boston, MA 02115, USA.

AparnaVidyasagar, Nancy A Wilson and ArjangDjamali**Heat shock protein 27 (HSP27): biomarker of disease and therapeutic target*.Nephrology Division, Department of Medicine, University of Wisconsin, H4/564 CSC, 600 Highland Avenue, Madison, WI, 53792, USA

Arcangela Gabriella Manente1, Giulia Pinton1, Daniela Tavian2, Gerardo Lopez-Rodas3, Elisa Brunelli1, Laura Moro1*Dipartimento di ScienzeChimiche, Alimentari, Farmaceutiche e Farmacologiche.*Coordinated Sumoylation and Ubiquitination Modulate EGF Induced EGR1 Expression and Stability*.University of Piemonte Orientale "A. Avogadro", Novara, Italy.

Bohren KM, Nadkarni V, Song JH, Gabbay KH, Owerbach D.A M55V polymorphism in a novel SUMO gene (SUMO-4) differentially activates heat shock transcription factors and is associated with susceptibility to type I diabetes mellitus. J Biol Chem. 2004 Jun 25; 279(26):27233-8. Epub 2004 Apr 29.

Bossi A, Lehner B (2009) Tissue *specificity and the human protein interaction network*.MolSystBiol 5: 260.

Brylinski M and Skolnick J. (2008) *A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.* ProcNatlAcadSci *USA* 105:129-34

Chang-Hoon Woo and Jun-ichi Abe* *SUMO: A post-translational modification with therapeutic potential?* Aab Cardiovascular Research Institute, University of Rochester School of Medicine and Dentistry, 601 Elmwood Avenue, Box CVRI, Rochester, NY 14642

Cobb MH, Goldsmith EJ. *How MAP kinases are regulated.* The Journal of biological chemistry 1995; 270:14843–14846. [PubMed: 7797459]

Dayangku, Fatiha,Pengiran,Burut& Anwar Borai. *Serum heat shock protein 27 antigen and antibody levels appear to be related to the macrovascular complications associated with insulin resistance: a pilot study*. Faculty of Health & Medical Sciences, University of Surrey, Guildford, Surrey, GU2 7XH, UK.

Esther Pilla1, Ulrike Möller1, Guido Sauer2, Francesca Mattiroli3, Frauke Melchior4 and Ruth Geiss-Friedlander1***A novel SUMO1-specific interacting motif in Dipeptidyl Peptidase 9*

*(DPP9) that is important for enzymatic regulation.*Department of Biochemistry I, Faculty of Medicine, Georg-August-University of Goettingen, Humboldtallee 23, 37073 Goettingen, Germany.

Grosdidier A, Zoete V, Michielin O. *SwissDock, a protein-small molecule docking web service based on EADock DSS.*Swiss Institute of Bioinformatics, Quartier Sorge, BâtimentGénopode, CH-1015 Lausanne, Switzerland.

Grosdidier A, Zoete V, Michielin O. *Fast docking using the CHARMM force field with EADock DSS.*Swiss Institute of Bioinformatics (SIB), Quartier Sorge, BâtimentGénopode, CH-1015 Lausanne, Switzerland.

Hormozdiari F, Salari R, Bafna V, SahinalpSC.*Protein-protein interaction network evaluation for identifying potential drug targets.*School of Computing Science, Simon Fraser University, Burnaby, Canada.

Kerscher O (2007) *SUMO junction—what's your function? New insights through SUMO interacting motifs.*EMBO reports 8:550–555

Lin D-Y et al. (2006) *Role of SUMO-Interacting Motif in Daxx SUMO Modification, Subnuclear Localization, and Repression of Sumoylated Transcription Factors.*Molecular Cell 24:341–354.

Matic I et al. (2010) *Site-Specific Identification of SUMO-2 Targets in Cells Reveals an Inverted SUMOylation Motif and a Hydrophobic Cluster SUMOylation Motif.*Molecular Cell 39:641–652.

Parcellier A, Schmitt E, Gurbuxani S, Seigneurin-Berny D, Pance A, Chantôme A, Plenchette S, Khochbin S, Solary E, Garrido C.*HSP27 is a ubiquitin-binding protein involved in I-kappaBalphaproteasomal degradation*.

Rajan S,Torres J, Thompson MS, Philipson LH. *SUMO downregulates GLP-1-stimulated cAMP generation and insulin secretion.* Univ. of Chicago, Chicago, IL 60637, USA.

Ramirez F, Schlicker A, Assenov Y, Lengauer T, Albrecht M (2007) *Computational analysis of human protein interaction networks.* Proteomics 7: 2541–2552.

Roberts PJ, Der CJ. *Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer.* Oncogene 2007;26:3291–3310.

Rosas-Acosta G, Russell WK, Deyrieux A, Russell DH, Wilson VG (2005) *A universal strategy for proteomic studies of SUMO and other ubiquitin-like modifiers.*Mol Cell Proteomics 4:56–72.

Skolnick J and Brylinski M. (2009),*FINDSITE: a combined evolution/structure-based approach to protein function prediction*. Brief Bioinform 10:378-91

Seo HR, Chung DY, Lee YJ, Lee DH, Kim JI, Bae S, Chung HY, Lee SJ, Jeoung D, Lee YS. *Heat Shock Protein 25 or Inducible Heat Shock Protein 70 Activates Heat Shock Factor 1 dehosphorylation on serine 307 through inhibition of ERK1/2 phosphorylation*. Laboratory of Radiation Effect and Laboratory of Radiation Experimental Therapeutics, Korea Institute of Radiological and Medical Sciences, Seoul 139-706

Tatham MH, Matic I, Mann M, Hay RT (2011) *Comparative Proteomic Analysis Identifies a Rolefor SUMO in Protein Quality Control.*Science Signaling4:rs4.

Umar-FaruqCajee, Rodney Hull and Monde Ntwasa **Modification by Ubiquitin-Like Proteins: Significance in Apoptosis and Autophagy Pathways*. School of Molecular & Cell Biology, Gatehouse 512, University of the Witwatersrand, Johannesburg, 2050

Vertegaal ACO et al. (2006) *Distinct and Overlapping Sets of SUMO-1 and SUMO-2 Target Proteins Revealed by Quantitative Proteomics*. Molecular & Cellular Proteomics 5:2298–2310.

YohannesMebratuand YohannesTesfaigzi.*How ERK1/2 Activation Controls Cell Proliferation and Cell Death Is Subcellular Localization the Answer?*Lovelace Respiratory Research Institute, Albuquerque, NM 87108, USA

Zhang C, Vasmatzis G, Cornette JL, DeLisi C. *Determination of atomic desolvation energies from the structures of crystallized proteins.* J Mol Biol. 267(3):707-26, 1997

Zhang X-D et al. (2008) *SUMO-2/3 Modification and Binding Regulate the Association of CENPE with Kinetochores and Progression through Mitosis.*Molecular Cell 29:729–741.

Zhu J et al. (2008) *Small ubiquitin-related modifier (SUMO) binding determines substrate recognition and paralog-selective SUMO modification.*J BiolChem283:29405–29415.

APPENDIX

OMIM:

OMIM is Online Mendelian Inheritance in Man is freely available web server which updated daily. The comprehensive briefing of genetic phenotypes and human genes takes place in OMIM. Information on all known Mendelian disorders and over 12,000 genes are present as full-text, referenced overviews in OMIM. The relationship between phenotype and genotype can be determined with the help of OMIM. Copious links to other genetics resources are found in the entries.

All the known diseases with a genetic component are archieved in OMIM. On human genome, these diseases are linked with the relevant genes. This provides references for further research and genomic analysis of an archieved gene.

Dr. Victor A. McKusick initiated this database in the early 1960s by archieving disorders and mendelian traits, entitled as Mendelian Inheritance in Man. Between 1966 and 1998, 12 books of MIM were published. In 1985, National Library of Medicine and the Medical Library at Johns Hopkins created the online version of OMIM which was made available on the internet in 1987. In 1995, NCBI developed OMIM for the World Wide Web.

ZINC Database

The ZINC database is a curated collection of commercially available chemical compounds, many of them “drug-like” or “lead-like” prepared especially for virtual screening. In research universities, biotech companies, and pharmaceutical companies, ZINC is used by investigators for identifying chemical compounds.

There are over 21 million purchasable compounds in ZINC which are used by many popular docking programs.

Searching ZINC:

Users may search ZINC based on several criteria. On the left hand side of the Web page limits on net charge and molecular weight may be specified. The registration codes, unique serial number may be specified on the bottom left page of ZINC database. The specified ZINC codes matched molecules will be found. The specification of constraint on the compound vendor also takes place. Using the Java Molecular Editor (JME).31, molecular substructures can be drawn.

KEGG (Kyoto Encyclopedia of Genes and Genomes)

KEGG deals with enzymatic pathways, biological chemicals and genomes. In 1995, Japanese human genome programme initiated the Kyoto Encyclopedia of Genes and Genomes.The networks of molecular interactions and their variants in the cells are recorded in PATHWAY database. The FTP of KEGG is no longer free since July 2011.

For browsing and retrieval of data and also for modelling and simulation, KEGG database can be used.

On molecular interaction networks, KEGG finds information, about pathways and complexes, genes and proteins generated by genome projects and about biochemical compounds and reactions. There are efforts in progress to add information regarding ortholog clusters into the knowledge of KEGG.

Five different databases are maintained by KEGG: KEGG Pathway, KEGG Ligand, KEGG BRITE, KEGG Atlas and KEGG Genes.

A number of KEGG Pathways are Cellular Processes, Metabolism, Human Diseases, Drug development, Genetic Information Processing and Environmental Information Processing.

Ligand Database includes different kinds of ligands: Drug Compound, Enzyme, Glycan, RPAIR and Reaction.

DrugBank:

The DrugBank database combines detailed chemical, pharmaceutical and pharmacological data of drug with comprehensive information of sequence, structure, and pathway of drug target. There are about 6811 drug entries in the database out of which 5080 are experimental drugs, 1528 are FDA-approved small drug molecules, 150 are FDA-approved peptide drugs and 87 are nutraceuticals. 4294 non-redundant protein sequences are also linked to these drug entries. There are more than 150 data field entries in each DrugCard out of which half of them tells about drug/chemical data and the other half tells about drug target or protein data.

Jmol:

Jmol is a freely available online tool based on Java viewer. It is used for viewing three-dimensional chemical structures like biomolecules, chemicals, crystals and materials. A variety of file types and output is readed with the help of quantum chemistry programs and animation with the help of quantum programs.

Lipinski filters:

Lipinski filters is a software tool designed and maintained by Supercomputing facility of IIT Delhi. This tool is mainly used for drawing a chemical or drug molecule online.

It checks whether a drug satisfies the five Lipinski rules or not.

Lipinski Rule of five:

The difference between drug like and non-drug like molecules can be identified by using Lipinski rule of 5. Molecules agreeing with 2 or more of the following rules helps in prediction of success or failure of drug.

- Molecular mass ≤ 500 Da
- Lipophilicity must be high
- Number of hydrogen bond donors must be less than 5
- Number of hydrogen bond receptors must be less than 10
- Range of molar refractivity must be between 40-130

These filters help in development of drugs. Late-stage preclinical and clinical failures can also be prevented by its use.

PyMOL:

PyMOL is a molecular visualization system sponsored by user. It works on an open-source foundation. Schrödinger distribute and maintains it.

For customization of 3-D images of biomolecules, PyMOL is as a leading software package. It has more than 20 representations and 600 settings which helps users in controlling it precisely and powerfully.

Over 30 different file formats can be interpreted by PyMOL (PDB files, volumetric electron density maps, multi-SDF files). Stunning 3-D images can be created from file formats by first-time and expert users with the help of graphical user interface of PyMOL. Images are then saved as Session file, which tells about position of object, atom colour, molecular state and representation, frames.

Features:

- View 3D structures of biomolecules.
- Artistically rendering of figures
- Dynamic animation of molecule
- PyMol geometry export
- 3D data is presented with AxPyMol

Image representations:

In about 20 different ways data can be represented using PyMOL. CPK-like view is provided by spheres, volumetric views is provided by surface and mesh , bond connectivity is represented by lines and sticks, and secondary structure and topology can be identify with the help of ribbon and cartoon.