

Computational Prediction of SUMOylation sites in proteins

*A thesis submitted as a
Requirement for the
Partial fulfillment of the
Degree of Doctor of Philosophy*

By

Yogendra Ramtirtha

Roll No. 20143311



Department of Biology,

Indian Institute of Science Education and Research, Pune - 411008

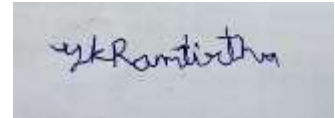
2022

To my parents

And brother

DECLARATION

I declare that this written submission represents my idea in my own words and where others' ideas have been included; I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.



Date: 09/12/2022

Yogendra Ramtirtha

20143311

IISER Pune

CERTIFICATE

Certified that the work incorporated in the thesis “Computational prediction of SUMOylation sites in proteins”, submitted by Yogendra Ramtirtha was carried out by the candidate under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other university or institution.

Date: 09/12/2022



Dr. M. S. Madhusudhan

Associate Professor

IISER Pune

ACKNOWLEDGEMENTS

The supervisor – My first interaction with Madhu took place in September 2013. Back in the day, Madhu was still at BII A-star in Singapore. I had a Skype call with him in order to join his lab for my Master's dissertation project. I started working in the lab in December 2013 and later joined the lab as Ph.D. student in August 2014. The lab moved its location 3 times during my tenure in the lab. It has been a long way. My dissertation project became my Ph.D. topic. I admire Madhu's critical thinking and data analysis skills. His ability to present his ideas to his audience with absolute clarity is another skill that I admire. Every relation has its ups and downs. Whenever the going gets tough, he would always keep his cool. His ambitious and idealistic nature is praise worthy. Hope to stay in touch in future.

Lab members – All the lab members that I have interacted with over the years have been very cooperative, welcoming and helpful. Neelesh, Neeladri, Sanjana, Sagar, Shipra, Abhilesh, Akash and Swastik were very helpful. Neelesh has excellent programming skills and helps very patiently when debugging complex code. Sagar suggested many useful ideas during lab meets. Neeladri made the lab a fun place. Sanjana is very hard working and she was the first person in the Pune lab to carry out experimental characterization of her findings. Parichit was a great scientific programmer and was the go to person when it came to HPC (Rosalind). KP, Minh and Binh were members of the Singapore lab who suggested many good suggestions during lab meets.

Later on, Nida, Anuj, Dayal, Ankit, Golding, Swastik and Gulzar joined the lab. Nida worked with Neeladri on development of substitution matrices. Anuj had

many useful tips when it came to dealing with different kinds of people. Dayal is a prodigy. Ankit and Golding were extremely diligent. Coffee breaks with Ankit and Golding were fun. I admire Ankit's logical thinking. It helped me in questioning many wrong assumptions I had made about my projects. I admire Golding's vocabulary and his presentation skills. Swastik was a hard worker too. Gulzar continued with Parichit's work and was good at web scraping. Shreyas was a post doc in the lab and was good at cheminformatics.

Tejashree, Atreyi and Mukundan joined the lab as juniors and are great Ph.D. students. Tejashree makes the lab a cheerful place. Atreyi is continuing Sanjana's wet lab experiments. Her lamin related work is interesting too. Mukundan is good at protein design as well as puzzle solving. I would like to thank Atreyi and Mukundan for proofreading the current thesis.

Samarpita, Devatrisha, Ashwin and Kaustubh did their summer and winter projects with me. It was fun to work with them. There were many rotation students whose names I could not include here. However, their contributions have not been forgotten.

Collaborators and RAC members – I would like to thank my RAC members Dr. Gayathri Pananghat and Dr. Nagaraj Balasubramanian for their helpful insights during my RAC meetings. I would like to thank Dr. Girish Ratnaparkhi and his lab members for their helpful suggestions. I would also like to convey my gratitude to Dr. Jomon Joseph from NCCS, Pune for the argonaute project, which gave me my first publication. I am thankful to Dr. Prasanna Venkatraman from ACTREC, Navi Mumbai for giving me a chance to work with 14-3-3 proteins and learn molecular dynamics simulations.

Administration – I would like to thank Dr. Girish and Dr. Srabanti (former and present) dean of doctoral studies. I would like to thank Tushar, Sayalee and Nayana from the admin office for helping me with paperwork. I would like to thank Dr. V.S.Rao for managing my fellowship. I would also like to mention Mrinalini and Mahesh from bio office for their help and cooperation. I would like to thank the Srinivasa Ramanujan library for providing access to scientific literature.

Family – My family (parents and brother) have been a pillar of strength throughout the course of my degree. There were many times when I wanted to give up but they always stood by me and boosted my morale. I am indebted to them for their motivation and support. My brother went a step further and even suggested frequent item-set mining (Apriori algorithm) which helped me a lot for data analysis and pattern finding. Whatever little success I have been able to achieve in my life, I owe it to my family.

Funding – Money is indispensable to any research project in general. I am grateful to DBT for providing me with fellowship. I would also like to convey my gratitude to Scivic Engineering Pvt Ltd and innoplexus Consulting Services Pvt Ltd for providing financial support during my extension years.

-Yogendra

Contents

Chapter 1: Overview	10
1.1 Synopsis.....	10
1.2 Thesis Organization.....	12
Chapter 2: A General Introduction to SUMOylation.....	14
2.1 The SUMO pathway	14
2.2 Experimental determination of SUMOylation sites	16
2.3 Computational approaches to predict SUMOylation sites	16
Chapter 3: Prediction of SUMOylation targets in <i>Drosophila melanogaster</i>	19
3.1 Introduction.....	19
3.2 Materials and Methods.....	21
3.2.1 Overview of the method used to identify fly orthologs	21
3.2.2 Reference database and query proteins	23
3.2.3 Frequent item-set mining based clustering of 15-mers centered on annotated SUMOylation sites	26
3.2.4 Hierarchical clustering of protein sequences	29
3.2.5 Clustering of Gene Ontology terms	31
3.2.6 Computational tools and programs used in this study.....	31
3.3 Results.....	31
3.3.1 Quantitative summary of results	31
3.3.1.1 Predictions from human protein PSIBLAST	31
3.3.1.2 Predictions from mouse protein PSIBLAST.....	35
3.3.2 Proteins identified by different fly SUMO proteomics experiments.....	38
3.3.3 Motifs obtained from clustering analysis of 15-mer centered on SUMOylation sites	40
3.3.4 Motifs obtained from clustering analysis of protein sequences	43
3.3.5 New predictions made using motifs obtained from protein sequence clusters.....	46
3.3.6 Analysis of Gene Ontology terms.....	48
3.3.7 Comparison between predictions from our study, GPS-SUMO and JASSA	55
3.4 Discussion.....	56
Chapter 4: SUMO-ON-THE-FLY web server.....	59
4.1 Introduction.....	59
4.2 Server description	59

4.3 Case study: 14-3-3 epsilon.....	60
4.4 Acknowledgements.....	61
Chapter 5: 3-D structure based prediction of SUMOylation sites	62
5.1 Introduction.....	62
5.2 Materials and methods	64
5.2.1 Generation of a dataset of SUMOylated protein structures	64
5.2.2 Computational tools used in this study	65
5.2.3 Sampling method to dock target proteins onto <i>ubc9</i>	65
5.2.3.1 Move target protein near <i>ubc9</i> active site.....	71
5.2.3.2 Optimize lysine torsion angles	74
5.2.3.3 Spin target protein around lysine.....	75
5.2.4 Discriminating poses based on residue contacts	77
5.2.5 Statistical parameters to assess predictions	78
5.3 Results.....	79
5.3.1 Analysis of the target protein structures in the dataset.....	79
5.3.2 Predictions made using residue contacts.....	84
5.4 Discussion.....	88
Chapter 6: Conclusions and Future prospects.....	90
Chapter 7: Publications	92
Chapter 8: Appendix	93
Chapter 9: Supporting Information.....	94
Chapter 10: Bibliography	95

Chapter 1: Overview

1.1 Synopsis

Proteins carry out various biological functions of a cell. These functions range from generating the energy currency of the cell also known as ATP (Adenosine Triphosphate), to maintaining the shape and structure of the cell, brought about by cytoskeletal elements. Amino acids are the building blocks of proteins. There are about 20 standard amino acids. Proteins are synthesized by the ribosomal machinery in a process referred to as translation. After translation, some proteins get covalently attached to different chemical moieties. Examples of such moieties include phosphate group (phosphorylation), acetyl group (acetylation), methyl group (methylation) etc. In some cases, the moiety could be a protein, for example ubiquitin (ubiquitination), SUMO (SUMOylation), nedd (neddylation) and other members of ubiquitin-like family of proteins.

A growing body of recent scientific literature has pointed to the importance of SUMOylation in regulating many cellular activities. SUMOylation is a post translational modification that involves formation of a covalent bond between C-terminus of a protein called SUMO (Small Ubiquitin-related MOdifier) and lysine residues in proteins from eukaryotic organisms. Experimental determination of SUMOylation sites is cumbersome due to many technical reasons. Owing to these reasons, computational methods for predicting SUMOylation sites are important. This thesis focuses on development of computational methods to predict SUMOylated lysines. The computational methods discussed in this thesis fall into two broad categories – sequence and structure-based.

The sequence-based method begins with a list of >9000 SUMOylated lysines from human and >900 mouse proteins. This list was obtained from recent mass spectrometry based proteomics experiments. The protein sequence alignment tool PSIBLAST was used to identify proteins from the fruit fly *Drosophila melanogaster* that were homologs of human and mouse SUMOylated proteins. Protein sequence alignments were scanned to identify lysine residues that were conserved between human / mouse as well as fly proteins. This method identified >8600 fly proteins encoded by >4600 fly genes as putative SUMOylation targets. The homology data was further analyzed to obtain three kinds of information. First analysis was carried out to identify amino acid residues that co-occur along with the conserved lysines. This helped in finding out sequence motifs involving the conserved lysines, for example ψ -K-x-(E/D), where ψ – I/L/V, K – SUMOylated lysine and x – any amino acid. Second analysis helped in identifying which protein families tend to get more SUMOylated than others, for example transcription factors such as zinc finger proteins. Third analysis helped in determining the preferred cellular localization, molecular function and biological activity of the proteins identified in this study. The results from this study will be made available to the scientific community in the form of a database called SUMO-ON-THE-FLY.

This thesis also presents a novel structure based method. The method was demonstrated with the help of a pilot / proof-of-concept study. The dataset for this study consisted of 1841 Protein Data Bank structures of known human SUMOylated proteins. A special docking tool referred to as “sampling method” was designed to model complexes between ubc9 and target (substrate) proteins. Ubc9 is the enzyme capable of distinguishing between SUMOylated and non-SUMOylated lysine residues. The ubc9-target complexes modeled using sampling

method, were analyzed in terms of the residue contacts at protein-protein interfaces. These residue contacts were further used to make predictions of SUMOylated and non-SUMOylated lysines. The structure based method proposed in this study achieved an accuracy of 81% and Matthews' Correlation Coefficient of 0.4.

1.2 Thesis Organization

Chapter 1: Overview

This chapter gives the synopsis for the present thesis and explains the organization of different chapters in it.

Chapter 2: A General Introduction to SUMOylation

This chapter gives a brief introduction to SUMOylation, its biological implications and experimental / computational approaches to study the modification.

Chapter 3: Prediction of SUMOylation targets in *Drosophila melanogaster*

This chapter presents a sequence based method to predict SUMOylation sites across different organisms. The method makes use of homology information. In addition, the method also presents three different kinds of clustering methods to obtain information about sequence motifs, protein families and biological functions of homologous proteins identified in this study.

Chapter 4: SUMO-ON-THE-FLY web server

The information obtained in the previous chapter from the homology based study will be presented to the scientific community in the form of a web server called SUMO-ON-THE-FLY.

Chapter 5: 3-D structure based prediction of SUMOylation sites in proteins

All the currently available SUMOylation site prediction tools are sequence based. This chapter presents a first of its kind structure based SUMOylation site prediction tool. The method presented in this chapter can overcome the drawbacks of all the sequence based tools. This tool achieved an accuracy of 81% and Matthew's correlation coefficient of 0.4.

Chapter 6: Conclusions and Future Prospects

This chapter highlights the contributions of this thesis to our better understanding of biology.

Chapter 7: Publications

This chapter gives a list of research articles that have already been published or those that will be published in near future (hopefully).

Chapter 8: Appendix

This chapter gives a summary of collaborative side projects.

Chapter 9: Supplementary Data

This chapter provides information about how to download supporting information of the present thesis.

Chapter 10: Bibliography

This chapter provides a list of all the references cited in this thesis.

Chapter 2: A General Introduction to SUMOylation

SUMOylation was discovered in the year 1996 as a post translational modification of the protein RanGAP1 (Ran GTPase Activating Protein-1) [1]. Ever since then, a lot of scientific research has been done to uncover biological implications of this modification. This chapter provides a brief overview of SUMOylation and different aspects associated with it. Each succeeding chapter in this thesis will have its own introduction. (Detailed reviews of SUMOylation can be found here [2–6]).

2.1 The SUMO pathway

SUMO (Small Ubiquitin-related MOdifier) is a protein of 90 – 100 amino acid residues. It is structurally similar to ubiquitin. Both the proteins belong to the beta grasp fold of proteins. Similar to ubiquitin, SUMO is conserved across all eukaryotes, from yeast to humans. Just like the ubiquitin pathway, the SUMO pathway consists of SUMO specific proteases, E1, E2 and E3 ligases (Figure 1) [2,3,5]. SUMO is synthesized in its precursor form in cells (Figure 1). The precursor undergoes proteolytic maturation at its C-terminus, carried out by SUMO specific proteases. The mature SUMO protein has a diglycine motif (-GG) at its C-terminus. The diglycine motif is conserved across all members of ubiquitin-like family of proteins such as ubiquitin, SUMO, NEDD and others.

The SUMO E1 activating enzyme (SAE1 / SAE2) uses energy from ATP (Adenosine Triphosphate) to form a thioester bond between C-terminus of mature SUMO and itself. Subsequently, SUMO E1 activating enzyme transfers the SUMO to SUMO E2 conjugating enzyme (UBC9). UBC9 has been shown to identify lysine residues in target (substrate) proteins that have a sequence motif ψ -K-x-(E/D), where ψ = Leu, ILE or Val, x is any amino acid and E/D is a Glu / Asp residue. UBC9 catalyzes the formation of covalent bond between C-terminus of SUMO and the side chain ϵ -amino group of lysine residues that conform to the

consensus motif. SUMO E3 ligase helps UBC9 to identify lysine residues that do not follow the consensus motif.

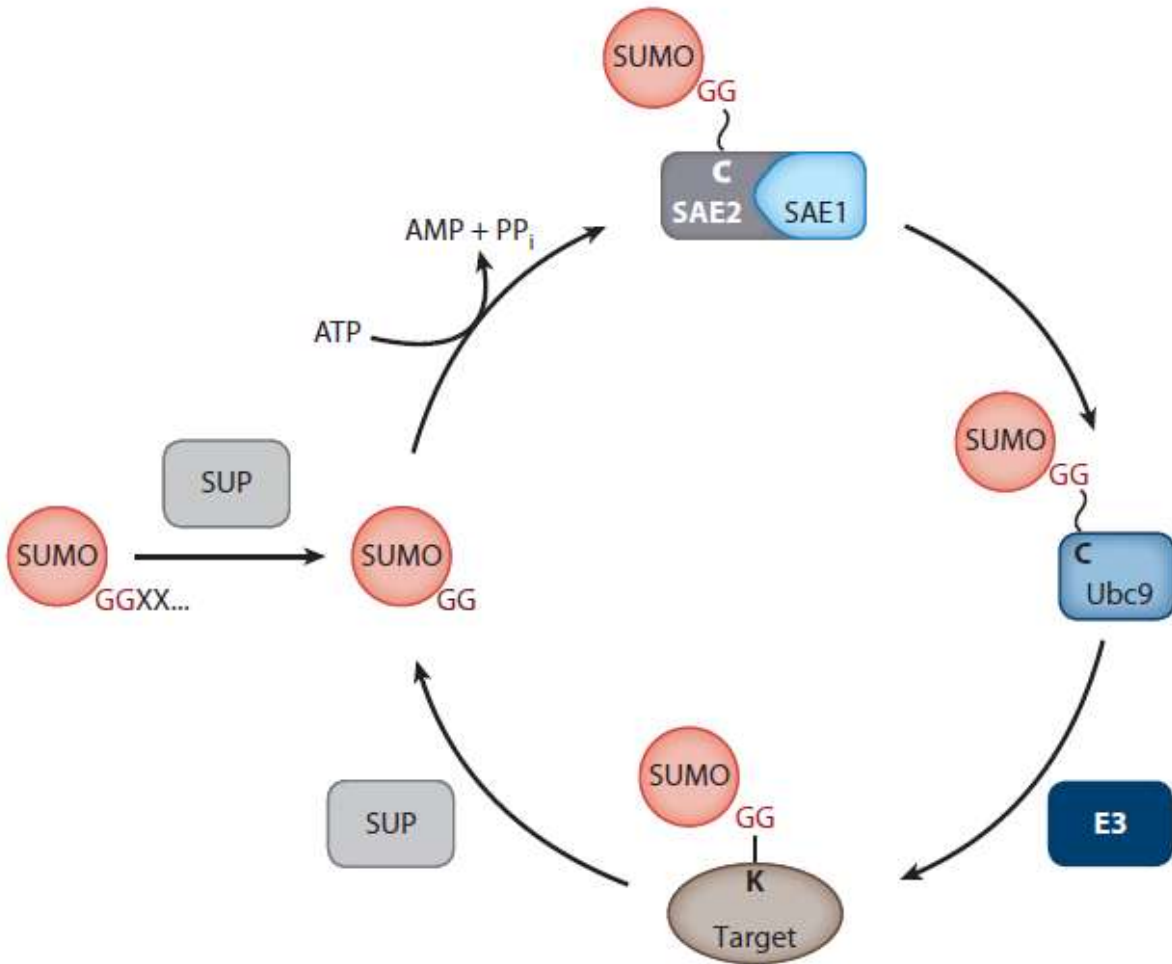


Figure 1: Schematic representation of the SUMO pathway. SUP refers to SUMO specific proteases. SAE1 / SAE2 refers to SUMO E1 activating enzyme. Ubc9 refers to SUMO E2 conjugating enzyme. E3 refers to SUMO E3 ligase. Adapted from [2].

From a cellular perspective, SUMOylation regulates most of the nuclear proteins, although it does regulate some of the cytosolic proteins as well [2,3]. Disruption of SUMOylation has been linked to various neurodegenerative diseases such as Alzheimer's, Huntington's, Parkinson's diseases among others [4]. Deregulation of SUMOylation has also been associated with cancer [5]. SUMOylation also controls antiviral innate immune response. Hence, viruses exploit host SUMOylation pathway in order to evade immune response [6].

2.2 Experimental determination of SUMOylation sites

Earlier SUMOylation studies involved lysine to arginine mutations of every lysine in a given protein. However, such studies were cumbersome for many technical reasons. These reasons include but are not limited to – 1: Activity of SUMO proteases (making detection difficult), 2: Low turnover of SUMOylation (a very small fraction of protein gets SUMOylated) and 3: Multiple lysines getting SUMOylated (hence need multiple lysine to arginine mutations). A comprehensive list of SUMOylation sites detected by these low throughput studies is maintained by many databases such as UniProt KB [7], dbPTM [8], PLMD [9] and PhosphoSitePlus [10].

In recent years, development of mass spectrometry based proteomics methods has enabled the detection of thousands of SUMOylation sites in a single experiment [11]. These experiments have shown that approximately 18% of the human proteome undergoes SUMOylation. SUMOylation is a very dynamic modification [11–13]. This means that the SUMOylome (SUMOylated fraction of proteome) depends on the cellular conditions. Thus, cells grown under standard growth conditions will have a different SUMOylome compared to cells grown under stress conditions such as heat shock or proteasome inhibitors.

2.3 Computational approaches to predict SUMOylation sites

The important role played by SUMOylation in regulating key biological processes has motivated researchers across the globe in developing many *in silico* prediction tools. Examples of popular prediction tools are GPS-SUMO [14,15] and JASSA [16]. Other examples of SUMOylation site prediction tools include SUMOpre [17], SUMOhydro [18], SUMOhunt [19], SUMOsu [20], SUMOgo [21], SumSec [22], HseSUMO [23] and pSUMO-CD [24]. (A detailed discussion of SUMOylation site prediction tools can be found here [25,26]).

All the presently available SUMOylation site prediction tools make use of information from protein sequences. The development pipeline of these tools can be summarized as follows.

- Literature available in PubMed and databases such as CPLM are searched to obtain a list of proteins and the lysines therein that get SUMOylated. The protein list is divided into training and testing datasets.
- Sequences of SUMOylated proteins are downloaded from databases such as UniProt. Information involving lysines is extracted in the form of peptides having size = k. For example, k = 15 for GPS-SUMO (7 residues before and after the lysine) and k = 21 for JASSA (10 residues before and after the lysine) respectively. Peptides centered on SUMOylated as well as non-SUMOylated lysines are extracted. Peptides from SUMOylated lysines are treated as positive dataset whereas peptides from non-SUMOylated lysines are treated as negative dataset.
- Information from k-mer peptides is encoded using feature extraction methods such as binary encoding, position specific scoring matrices and hydrophobicity scales.
- Proportions of positive and negative datasets are balanced using under-sampling or over-sampling. From all the available features, only those are retained that have statistically significant information.
- Prediction tools are developed by training machine learning algorithms such as support vector machines, artificial neural networks, K nearest neighbors and random forests.
- Statistical measures such as sensitivity, specificity, accuracy and correlation coefficients are used for performance assessment.

Despite all the efforts dedicated to the development of existing SUMOylation site prediction tools, these tools have drawbacks. Most important drawback of these tools is that they predict a lot of false positives. Such false positive predictions are futile because they do not provide any meaningful biological information. Another drawback affecting existing computational methods is that they over-predict lysine residues following the consensus motif ψ -K-x-(E/D) and under-predict lysines not following the motif. This bias hinders the accuracy of existing computational methods. SUMOylation has been known to happen to both kinds of lysine residues – those that follow the consensus motif as well as those that do not follow the motif.

In order to overcome the shortcomings of existing tools, the present thesis proposes two novel computational methods. Chapter 3 describes a method that uses protein sequences to extract evolutionary information, which in turn is used to predict SUMOylated lysines conserved across different organisms. Chapter 5 of this thesis describes a method that utilizes protein 3-D structures to differentiate between lysine residues that can bind active site of the enzyme ubc9 from those lysines that cannot.

Chapter 3: Prediction of SUMOylation targets in *Drosophila melanogaster*

3.1 Introduction

Small Ubiquitin-related MOdifier (SUMO) is a post-translational modifier protein of ~100 amino acids conserved in all eukaryotes from yeast to humans. SUMO is structurally similar to ubiquitin. Both the proteins belong to the beta-grasp fold of proteins. The SUMO pathway contains SUP (SUMO-specific Proteases), E1, E2 and E3 enzymes. Covalent attachment of SUMO C-terminus to NZ atom of lysine residue in target proteins is known as SUMOylation. SUMOylation has been shown to regulate various biological processes such as nucleo-cytoplasmic translocation of proteins and the activity of transcription factors. Disruption of SUMOylation has been linked to various neuro-degenerative diseases and cancer (a detailed review of SUMOylation can be seen in [2]).

In the last few years, development of mass-spectrometry coupled proteomics experiments has enabled identification of SUMOylation sites in thousands of human proteins (a detailed review can be found in [11]). In addition to standard cellular growth conditions, these experiments also probed the SUMOylation status of proteomes from cells grown under different stress conditions such as heat shock and proteasome inhibitors. Recently, [13] have come up with a list of human and mouse SUMOylated lysines and proteins using mass-spectrometry based proteomics approach.

SUMO-proteomics experiments in the fruit fly *Drosophila melanogaster* have studied the cellular effects of the modification [27–29]. These experiments

identified SUMOylated proteins but technical difficulties hindered the identification of modified lysines in these proteins. Knowledge of SUMOylated lysines is important for understanding biological implications of the modification. There are 2 computational approaches to predict putative SUMOylation sites. First method involves using currently available SUMOylation site prediction tools such as GPS-SUMO [30,31] and JASSA [16]. These tools use protein sequence as input and scan local sequence environment around lysine residues to make predictions. However, these tools miss information about known SUMOylated lysines from homologous proteins. Second method to predict putative SUMOylation sites is presented in this study. The proposed method begins by identifying fruit fly homologs of known SUMOylated proteins from other organisms. Homology search is based on protein sequence alignments. Putative SUMOylation sites are annotated based on quality of sequence alignments and local sequence environment around lysine residues.

In this study, human and mouse SUMOylated proteins were used to identify orthologous proteins from the proteome of fruit fly *Drosophila melanogaster*. Homology search was carried out using the sequence alignment tool PSIBLAST (Position Specific Iterative Basic Local Alignment Search Tool). In addition, sequence patterns involving SUMOylation sites were studied. Two kinds of information were obtained from sequence pattern analysis. First kind of information was obtained by analyzing local sequence around SUMOylated lysines to detect new motifs. Second kind of information was extracted by clustering target proteins according to their families and identifying lysines conserved in every family. Apart from sequence patterns, Gene Ontology analysis was also carried out to study preferred cellular compartments and biological processes of the modified proteins.

3.2 Materials and Methods

3.2.1 Overview of the method used to identify fly orthologs

The objective of this work was to identify fly homologs of human and mouse SUMOylated proteins. For this reason, human and mouse proteins were queried against a reference database containing the fruit fly proteome and UniRef90 proteins using PSIBLAST. All the details concerning the list of human proteins, reference database and PSIBLAST parameters will be discussed in the following subsections. Alignments involving fly-human and fly–mouse protein pairs were extracted from PSIBLAST results. All the pair-wise alignments were scanned to check whether SUMOylated lysines from human proteins were aligned to a lysine residue from the corresponding fly protein. If this was the case, then 15-mers centered on the lysines of interest were extracted. The 15-mers were extracted by including 7 residues upstream and downstream with respect to lysine of interest. There were 2 lists of 15-mers per organism. The first list was extracted from the FASTA protein sequences (referred to as FASTA 15-mers). And the other list was extracted from the aligned region of the protein sequence (referred to as Aligned 15-mers, which may contain gaps from alignments). The workflow for obtaining 15-mers is summarized below (Figure-1).

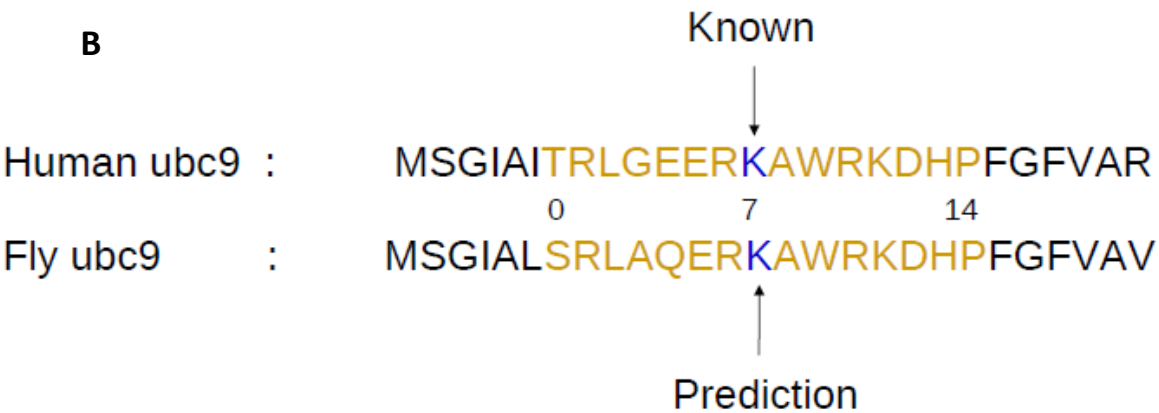
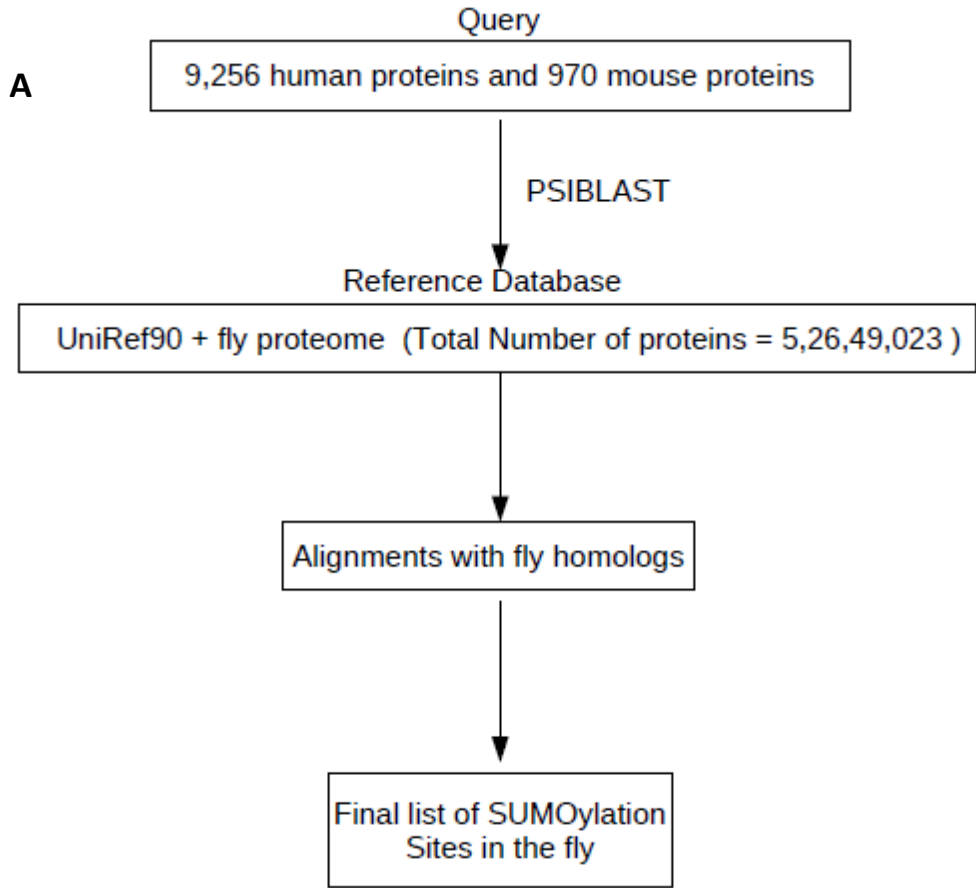


Figure 1: **A:** Overview of methods used to identify fly orthologs and 15-mers centered on aligned SUMOylated lysines using human and mouse proteins with known SUMOylated lysines. The homology search was carried out using PSIBLAST. **B:** An example alignment between human and fly homologs of the SUMO E2 conjugating enzyme (ubc9) and the conserved lysines as well as 15-mers therein. The alignment shown here has an E-value of 1e-78.

3.2.2 Reference database and query proteins

For this study, the list of human and mouse SUMOylated proteins was obtained from supplementary data of [13] (human data file named - 41467_2018_4957_MOESM6_ESM.xlsx and mouse data file named - 41467_2018_4957_MOESM8_ESM.xlsx) . Details of query proteins used in this study are summarized below (Table-1).

Table-1: Details of query proteins used in this study

Query type	Number of proteins	Source Organism	Reference
1	9256	Human	[13]
3	964	Mouse	[13]

The FASTA sequences of all the query proteins and reference databases were downloaded from the UniProt database (<https://www.uniprot.org/>) [7].

The reference database (5,26,49,023 proteins) used in this study was created by combining UniRef90 database (5,26,29,880 proteins) and fruit fly proteome (19,143 proteins, Tax ID – 7227). In order to remove duplicates in the reference database, protein sequences in the UniRef90 database that came from the fruit fly proteome (TaxID – 7227) and having the term “Drosophila melanogaster” in their header were removed.

Homologs of query proteins were searched in the reference proteome using PSIBLAST version 2.7.1+ [32,33]. For every query protein, 2 kinds of PSIBLAST

jobs were carried out. The parameters used for both kinds of jobs are summarized below (Table-2).

Table-2: Parameters of both PSIBLAST jobs

Parameter	Job1	Job2
Number of rounds	5	5
E-value cutoff to include in PSSMs	10^{-5}	10^{-5}
Number of alignments	1000	100000
Seg (mask low complexity region in query)	Yes	Yes
Composition based statistics	0	0

For every query protein, 2 kinds of PSIBLAST jobs were carried out. The PSSMs (Position Specific Scoring Matrices) generated after both the jobs differ because of the difference in the number of alignments used namely – 1,000 and 100,000 respectively. The job with 1,000 alignments was carried out to detect close homologs, whereas the job with 1,00,000 alignments was carried out to detect distant homologs. Pair-wise alignments in different rounds of PSIBLAST use different PSSMs. Thus, E-values from different rounds for the same protein pair cannot be compared with each other because PSIBLAST takes PSSMs into account while calculating E-values. Hence, E-values for all pair-wise alignments between human – fly and mouse – fly proteins were re-calculated using BLOSUM62 substitution matrix and the E-value equation (Equation 1).

Equation 1

$$E = mn2^{-S'}$$

Equation 2

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

Where E = E-value, S = raw score calculated using BLOSUM62 substitution matrix, S' = bit score, K and λ are constants, $\ln K$ = natural logarithm of K and $\ln 2$ = natural logarithm of 2. In addition to substitution scores, raw score calculation also took into account affine gap penalties with a gap existence penalty = 11 and gap extension penalty = 1.

For this study, $K = 0.041$, $\lambda = 0.267$, m = length of query protein and total number of amino acids in fly proteome also known as n = entire length of fly proteome present in the reference database = 17,879,049,827, were used for E-value calculations. All the equations and parameters used here were taken from PSIBLAST results. For scoring 15-mer alignment, m = 15 and n = 8,629,350 were used. The value of n was calculated after taking into account all possible 15-mers centered on all lysines of fly proteome present in the reference database.

For every protein pair (namely human-fly and mouse-fly), there are 10 alignments from PSIBLAST (both the jobs yield 5 alignments each) results, including all the alignments that would have introduced bias in clustering analysis discussed later. Hence, for every protein pair the alignment with lowest re-calculated E-value was chosen for further analysis.

3.2.3 Frequent item-set mining based clustering of 15-mers centered on annotated SUMOylation sites

Frequent item-set mining methods are commonly used for finding patterns in customer transaction data in the retail industry, for example market basket analysis. Apriori algorithm is a commonly used method in the field of frequent item-set mining. In this study, Apriori algorithm [34] was used to detect commonly occurring amino acid patterns in the 15-mer data obtained from PSIBLAST analysis.

The human and mouse query proteins as well as the fly homologs identified in the respective PSIBLAST jobs have redundancy. This could introduce bias in the clustering analysis. Hence, all of these protein lists were culled using h-CD-HIT server (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=h-cd-hit) [35–38]. The culling process was hierarchical in nature and was carried out with 3 different identity cutoffs – 90%, 60% and 30% respectively (the results discussed here were obtained at 30% redundancy). Details of non-redundant protein lists and 15-mers centered on SUMOylation sites present in these proteins are given below (Table-3).

Table-3: Summary of protein lists obtained from h-CD-HIT server

Query type	Protein type	Total list	nr (Non redundant list)	FASTA 15-mers (nr)	Aligned 15-mers (nr)
Human	Human	5283	3170	10245	10245
Human	Fly	8539	3725	8657	8657
Mouse	Mouse	468	373	549	549
Mouse	Fly	1700	707	876	876

It should be noted that for every fly homolog, only those 15-mers were chosen that occurred in longest alignments for the given fly protein. This was done because some fly proteins are aligned to multiple query proteins. Including 15-mers from all the alignments would have introduced repetitions that would have affected the clustering analysis described below.

At this point, it is important to define terminologies used in this analysis. An item is the frequency of each of the 20 amino acids to occur at every position in the 15-mer. Positions in a 15-mer are numbered from 0 to 14, the SUMOylated lysine is at 7th position. The 20 standard amino acids are sorted in an alphabetical order of their one letter code starting with Alanine (A) and ending with Tyrosine (Y). The SUMOylated lysine (7K) is omitted from the analysis since it is common to all 15-mers. Possible items could be 0A, 0C, 0D 6V, 6W, 6Y and 8A, 8C, 8D 14V, 14W, 14Y. In other words, there are $14 \times 20 = 280$ unique items in the data of a given 15-mer category. Another term used in this analysis is called support, which is frequency of a given item (or item-set) in the data of a given 15-mer category divided by total number of 15-mers of a given category. In other words, support is probability or normalized frequency. All 15-mers having gaps or occurring near N-terminus or C-terminus are excluded from this analysis.

The Apriori algorithm can be explained in the steps given below –

- i. Calculate support values for each of the 280 items. Each item could be thought of as an item-set of size equal to 1. Select all size-1 item-sets that have their support greater than or equal to their support cutoff.

- ii. Generate all possible new item-sets of size = 2. New item-sets are generated by extending item-sets from previous step by an item, such that the item is a member of an item-set from previous step.
- iii. Support values are calculated for new item-sets. All item-sets with support greater than or equal to support cutoff are selected.
- iv. This process of generation and selection of new item-sets having size greater by 1 than previous step, is continued till no new item-sets can be created. At this step, the algorithm ends.

The item-sets having size-2 resulting from the Apriori algorithm were processed further. In case a group of size-2 item-sets satisfy the following conditions, they are combined together and their support values are added up –

- i. The given group of size-2 item-sets should have 1 common item and the varying item should have the same position but different amino acid.
- ii. All the item-sets of the given group should have their support values greater than or equal to support cutoff for combining 2-mer item-set for the given 15-mer category in accordance to Table-4.

Let us understand the combining exercise with an illustration. For example, consider 3 size-2 item-sets: 6I-9E, 6L-9E and 6V-9E. These item-sets have 1 common item namely 9E and the varying items – 6I, 6L and 6V have same position – 6 but varying amino acids namely – I, L and V. Since, the 2 conditions for combining size-2 item-sets has been met, the 3 item-sets will be combined into a consensus motif - 6[IVL]-9E, where I, V and L are the 3 amino acids that could occur at position 6 while E occurs at position 9. When the item-sets are combined into a consensus motif, their respective support values are added up and the sum is

assigned to the consensus motif. Given below are support cutoffs used for different 15-mer categories (Table-4).

Table-4: Support cutoffs for each 15-mer category used as input to Apriori algorithm and support cutoff for clustering 2-item-sets for each 15-mer category

Query type	Protein type	Support cutoff Apriori (FASTA 15-mer) %	Support cutoff to cluster 2-mer (FASTA 15-mer) %	Support cutoff Apriori (Aligned 15-mer) %	Support cutoff to cluster 2-mer (Aligned 15-mer) %
Human	Human	0.425	1.5	0.425	1.3
Human	Fly	0.65	0.82	0.65	0.79
Mouse	Mouse	2.6	3	2.6	3
Mouse	Fly	0.85	1.5	0.85	1.5

3.2.4 Hierarchical clustering of protein sequences

The results from h-CD-HIT server discussed in the previous section contain proteins clustered according to their sequence identities. Given below are details of clusters from each 15-mer category (Table-5).

Table-5: The Protein sequence clusters from h-CD-HIT server.

Query type	Protein type	Total number of clusters	Number of clusters size = 1	Number of clusters size ≥ 2	Number of proteins present in clusters size ≥ 2
Human	Human	3170	2180	990	3103
Human	Fly	3725	1876	1849	6663
Mouse	Mouse	373	309	64	159
Mouse	Fly	707	336	371	1364

Multiple sequence alignments were constructed for every cluster using SALIGN function in MODELLER 9.17 [39]. All SUMOylation sites that line up at the same position in the MSA were grouped together and 15-mers centered on these lysines were extracted (discussed later in Results section). These clusters of SUMOylation sites were sorted in descending order by their sizes. Motifs were extracted from these 15-mer clusters. In this case, a motif is a 15-mer sequence such that each of the 15 positions is represented by the most abundant amino acid at that position in the 15-mer cluster. In case a position has its most abundant amino acid having frequency less than 70%, then x is included at that position in the motif sequence. The frequency cutoff of 70% was used because any cutoff less than that would have been too lenient and it would have introduced noise in the motif sequence.

3.2.5 Clustering of Gene Ontology terms

This analysis was done to identify patterns related to protein functions for the identified homologs. Gene Ontology terms of every protein from both human and mouse PSIBLAST data were searched in the UniProt KB database. These terms can be divided into 3 categories - Cellular Component (GO C), Molecular Function (GO F) and Biological Process (GO P). It is possible for a protein to have one or more GO C, F and P terms associated with it. Each category of terms was clustered independently. All proteins that have the same term were grouped into the same cluster. For example, all proteins having Gene Ontology cellular component term as “nucleus” were grouped together into one cluster. Finally, all the clusters of a given category were sorted in descending order by their size.

3.2.6 Computational tools and programs used in this study

All the data extraction and analysis steps were carried out in Python version 2.7.5 and mathematical calculations were done using Numeric Python (NumPy) version 1.7.1 [40] respectively. Venn diagrams discussed later in the results section were plotted using VennDiagram library version 1.6.20 [41] in R version 3.4.4 [42] .

3.3 Results

3.3.1 Quantitative summary of results

3.3.1.1 Predictions from human protein PSIBLAST

Given below is a summary of the results obtained from PSIBLAST-based analysis of 9256 human proteins described in [13] (Table-6). Out of a total 9256 proteins, about 5283 (57%) proteins have around 8539 fly orthologs. As for the 15-mers

centered on SUMOylated lysines, 5283 human proteins contain 19823 15-mers centered on FASTA and aligned protein sequences. There around 52332 15-mers entered on annotated SUMOylated lysines from 8539 fly orthologs. There are about 4591 fly genes that encode for the 8539 fly proteins identified in this study. The CG-names and FlyBase FBgn identifiers for these genes were obtained from FlyBase release FB2018_05 (<https://flybase.org/>) [43] .

Table-6: Summary of results obtained from PSIBLAST of human proteins

Description	Numbers
Total number of human proteins	9256
Number of human proteins that found fly orthologs	5283
Number of fly orthologs found	8539
Number of aligned and FASTA human 15-mers	19823
Number of aligned and FASTA fly 15-mers	52332
Number of fly genes that encode the fly proteins	4591

SUMOylated lysines are known to conform to a consensus sequence motif – ψ – K – x – (E/D) – where ψ = any aliphatic, hydrophobic amino acid such as I / V / L, K = SUMOylation site, x = any amino acid and E/D = either glutamate or aspartate residue. Since ψ and x are variable amino acids, only the K-x-E motif was checked in the FASTA 15-mers centered on SUMOylation sites. Given below is a summary

of the proportion of human and fly SUMOylation sites that conform to the K-x-(E/D) motif (Table-7). The proportions of human and fly SUMOylation sites that do not conform to the consensus motif are 67% and 74% respectively.

Table-7: Summary of proportion of SUMOylation sites in human data that conform to K-x-(E/D) motif

Motif status	Proportion of human SUMOylation sites	Proportion of fly SUMOylation sites
K - x - (E/D)	3128 (16 %)	6292 (12 %)
(E/D) - x - K	2820 (14 %)	6225 (12 %)
(E/D) - x - K - x - (E/D)	517 (3 %)	1079 (2 %)
None	13358 (67 %)	38736 (74 %)

A Venn diagram comparing the overlap between lists of CG-names for genes encoding fly SUMO targets identified by different SUMO proteomics studies can be seen below (Figure-2). These studies could identify fly SUMO targets but not the SUMOylated lysines due to experimental difficulties. The CG-name lists were derived from fly proteins identified by human PSIBLAST data, Nie et al 2009 [27], Handu et al 2015 [28], Pirone et al 2017 (S2R+ cell lines) and Pirone et al 2017 (transgenic flies) [29] respectively. Around 1069 (23 %) of CG-names for fly orthologs found from the human PSIBLAST data were confirmed by at least one of the other studies (Figure-2). These 1069 CG-names also account for 91% of the 1169 CG-names identified by 2 or more studies compared in the Venn diagram given below (Figure-2). There are 5 proteins common among all studies. These 5 proteins are actin-5C, Hsp68, RNP-107kd, 14-3-3 epsilon and 14-3-3 zeta.

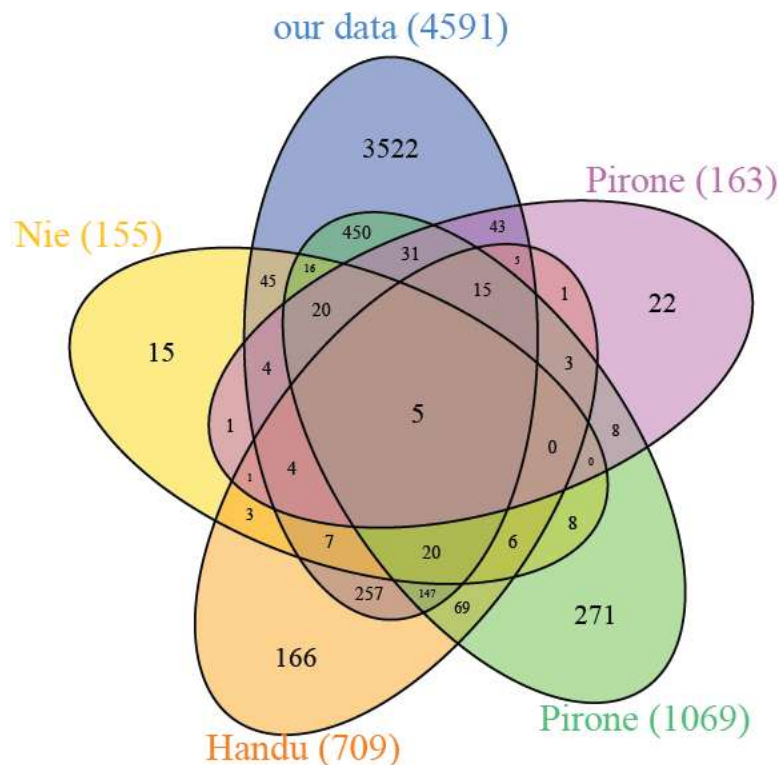


Figure 2: Venn diagram comparing gene CG-name list of fly orthologs from human PSIBLAST data with gene CG-name lists for fly SUMOylated proteins identified by other SUMO proteomics studies. Here, numbers in brackets indicate total number of CG-names for a given study and Nie – Nie et al 2009, Handu – Handu et al 2015, Pirone (1069) – Pirone et al 2017 data from S2R+ cell lines, Pirone (163) – Pirone et al 2017 data from transgenic flies.

Predictions from human PSIBLAST data also contain 3522 CG-names that were not identified by any of the 3 fly SUMO proteomics studies. 2194 of the 3522 fly genes encode at least one protein containing either K-x-(E/D), (E/D)-x-K or (E/D)-x-K-x-(E/D) consensus motifs. 1033 of the 3522 fly genes also encode proteins that have nucleus as their preferred cellular localization as indicated by their Gene Ontology Cellular Component terms. This observation is consistent with previous reports suggesting nucleus as a preferred cellular compartment of SUMOylated

proteins [11,13]. A detailed discussion of various Gene Ontology terms can be found later in this article.

3.3.1.2 Predictions from mouse protein PSIBLAST

Given below is a summary of the results obtained from PSIBLAST of mouse proteins (Table-8). Out of 970 mouse proteins, around 468 proteins (48%) have 1700 fly orthologs. There are 769 SUMOylated lysines in 468 mouse proteins. Similarly, the 1700 proteins contain 3700 annotated SUMOylated lysines. There are 936 fly genes that encode for the 1700 fly orthologs.

Table-8: Summary of results obtained from PSIBLAST of mouse proteins

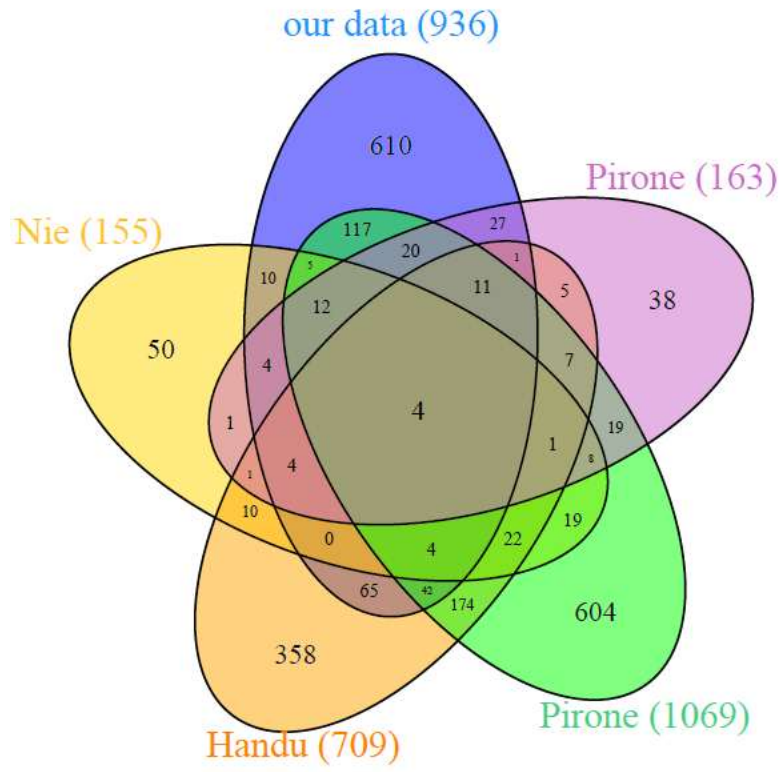
Description	Numbers
Total number of mouse proteins	970
Number of mouse proteins that found fly orthologs	468
Number of fly orthologs found	1700
Number of aligned and FASTA mouse 15-mers	769
Number of aligned and FASTA fly 15-mers	3700
Number of fly genes that encode the fly proteins	936

The proportions of K-x-(E/D) motif lysines in mouse and fly FASTA 15-mers are similar to the proportions reported for human PSIBLAST data in previous section (Table-9). Thus, around 57% of mouse SUMOylation sites and 76% of fly annotated SUMOylation sites do not conform to the K-x-(E/D) motif.

Table-9: Summary of proportion of SUMOylation sites in mouse data that conform to K-x-(E/D) motif

Motif status	Proportion of mouse SUMOylation sites	Proportion of fly SUMOylation sites
K – x – (E/D)	239 (31 %)	478 (13 %)
(E/D) – x – K	67 (9 %)	341 (9 %)
(E/D) – x – K – x – (E/D)	27 (3 %)	74 (2 %)
None	436 (57 %)	2807 (76 %)

A



B

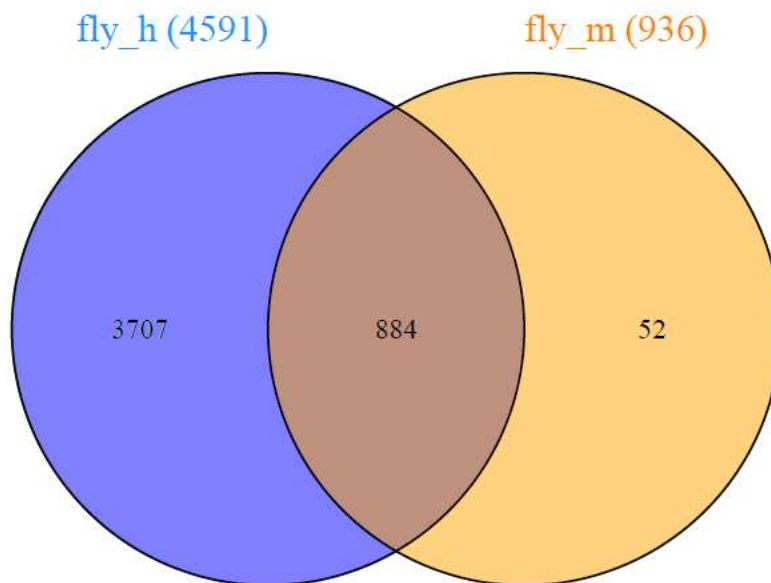


Figure 3: **A:** Venn diagram comparing gene CG-name list of fly orthologs identified from mouse PSIBLAST data with gene CG-name lists of fly SUMOylated proteins identified by other SUMO proteomics studies namely – Nie et al 2009, Handu et al 2015, Pirone et al 2017 (S2R+ cell lines) and Pirone et al 2017 (transgenic flies) Notations are same as Figure-2. **B:** Venn diagram comparing CG-name lists of fly orthologs identified by human PSIBLAST data and mouse PSIBLAST data.

Given above is a Venn diagram comparing overlap between CG-name lists of genes encoding fly orthologs identified from mouse PSIBLAST analysis and other SUMO proteomics studies (Figure-3A, Notations are the same as Figure-2). Around 326 (35%) of CG-names for fly orthologs found from mouse PSIBLAST analysis were confirmed by at least one of the other studies. The 4 proteins common in all studies are RNP-107kd, actin-5C, 14-3-3 epsilon and 14-3-3 zeta. In addition, overlap between CG-name lists for genes encoding fly orthologs identified by human PSIBLAST data and mouse PSIBLAST data can be seen above (Figure-3B). Around 884 (94%) of CG-names from mouse data were confirmed by CG-names from human data.

3.3.2 Proteins identified by different fly SUMO proteomics experiments

The Venn diagram derived from human PSIBLAST data (Figure-2) consists of proteins that belong to different overlap categories. Overlap categories could range from 0 to 4, where either a protein was detected by none of the fly proteomics studies to as many as all 4 studies. Proteins detected by all 4 studies have been discussed in the previous section. For the sake of analysis, proteins could be divided into 4 different categories, 1 – detected by 3 of the 4 studies, 2 – detected by 2 of the 4 studies, 3 – detected by 1 of the 4 studies and 4 – detected uniquely in this study. All the 4 categories are a subset of human PSIBLAST data. For each of these 4 categories, proteins were sorted in an ascending order according to the minimum E-value the given protein could achieve out of all the alignments

involving the given protein. Given below are top 5 proteins from every category (Table-10).

Table-10: Top 5 proteins and their respective UniProt / UniRef90 IDs for every overlap category.

Ran k	Proteins found by 3 studies	Proteins found by 2 studies	Proteins found by 1 study	Proteins unique to our study
1	P0CG69 : Polyubiquitin	P17210 : Kinesin heavy chain	A1ZAB5 : Protein clueless	X2JCN4 : Kugelei, isoform E
2	Q9TVM2 : Exportin-1	A4V1F9 : Ubiquitin- 63E, isoform C	A0A0B4KEX 0 : Protein clueless	P54358 : DNA polymerase delta catalytic subunit
3	P13060 : Elongation factor 2	Q59E34 : Mi- 2, isoform B	Q9VPI9 : LD20667p	Q7KU24 : Chromodomain -helicase- DNA-binding protein 1
4	P15348 : DNA topoisomeras e 2	E1JI46 : Mi-2, isoform C	M9PIA6 : Mi- 2, isoform D	UniRef90_Q8S WV9 : LD39323p (Fragment)
5	P11147 : Heat shock	Q9VF02 : Helicase 89B,	Q00174 : Laminin	P24014 : Protein slit

	70 kDa protein cognate 4	isoform B	subunit alpha	
--	--------------------------------	-----------	---------------	--

3.3.3 Motifs obtained from clustering analysis of 15-mer centered on SUMOylation sites

Shown below are top 5 motifs (in terms of support values) found in human PSIBLAST data using Apriori algorithm (Tables-11 and 12).

Table-11: Top 5 motifs found in FASTA and aligned 15-mers found in human proteins in human PSIBLAST data. Numbers in brackets indicate total number of motifs for the given category.

Rank	Motif human FASTA 15- mer (5882)	Support human FASTA 15- mer (%)	Motif human aligned 15- mer (5885)	Support human aligned 15- mer (%)
1	6[ILV]-9E	5.721	6[ILV]-9E	5.621
2	9E-11[EL]	3.538	9E-13[EKL]	4.577
3	8E-9E	2.013	2[EKL]-9E	4.017
4	9E-12E	1.904	9E-12[EK]	3.26
5	4E-5E	1.874	3[EL]-9E	3.026

Table-12: Top 5 motifs found in FASTA and aligned 15-mers found in fly proteins in human PSIBLAST data. Numbers in brackets indicate total number of motifs for the given category.

Rank	Motif fly FASTA 15-mer (832)	Support fly FASTA 15-mer (%)	Motif fly aligned 15-mer (889)	Support fly aligned 15-mer (%)
1	6[ILV]-9E	2.994	4[AEKLRS]-6L	5.272
2	9E-13[EKL]	2.982	8L-13[AKLR]	3.896
3	5[DEL]-6L	2.829	6[AILV]-9E	3.78
4	0[EKL]-8L	2.805	9E-13[AEKL]	3.78
5	4[AER]-6L	2.793	0[EKLR]-8L	3.737

As can be seen above, the most commonly occurring motif in human PIBLAST data is 6[IVL]-9E (Tables-11 and 12).

Given below are top 5 motifs (in terms of support values) found in mouse PSIBLAST data using Apriori algorithm (Tables-13 and 14).

Table-13: Top 5 motifs found in FASTA and aligned 15-mers found in mouse proteins in mouse PSIBLAST data. Numbers in brackets indicate total number of motifs for the given category.

Rank	Motif	Support	Motif mouse	Support
-------------	--------------	----------------	--------------------	----------------

	mouse FASTA 15- mer (39)	mouse FASTA 15- mer (%)	aligned 15- mer (28)	mouse aligned 15- mer (%)
1	6[ILV]-9E	21.642	6[ILV]-9E	19.744
2	4[EKSV]-9E	13.993	9E- 12[DESTV]	16.667
3	9E-13[EGP]	13.06	4[AEKS]-9E	14.103
4	9E-12[DES]	12.5	9E-13[EGP]	13.333
5	3[LPS]-9E	11.754	3[LST]-9E	11.538

Table-14: Top 5 motifs found in FASTA and aligned 15-mers found in fly proteins in mouse PSIBLAST data. Numbers in brackets indicate total number of motifs for the given category.

Rank	Motif fly FASTA 15- mer (838)	Support fly FASTA 15- mer (%)	Motif fly aligned 15- mer (1181)	Support fly aligned 15- mer (%)
1	6[IL]-9E	3.833	6[IL]-9E	3.406
2	5E-9E	1.974	8[PT]-12S	3.251
3	9E-10P	1.974	0Y-2C-5C- 9F	2.012
4	9E-11E	1.858	3E-4L	2.012
5	10L-14L	1.858	9E-11E	2.012

As can be seen above, 6[IVL]-9E is the amino acid motif with highest frequency in mouse PSIBLAST data (Tables-13 and 14).

All the motifs shown above (Tables-11 to 14) also contain the item 7K or central SUMOylated lysine. Recall from methods section that this item was omitted from input to Apriori algorithm because it occurs in all 15-mers. The item 7K should be considered while analyzing all the motifs (Tables-11 to 14). Thus, the motif 6[IVL]-9E can be expanded as 6[IVL]-7K-9E motif. In other words, this motif represents the SUMOylation consensus motif - $\psi - K - x - (E/D)$ - where ψ is an aliphatic hydrophobic amino acid I, V or L occurring at 6th position, 7K is the SUMOylated lysine and 9E indicates the glutamate residue occurring at the 9th position in the 15-mer sequence.

3.3.4 Motifs obtained from clustering analysis of protein sequences

Given below are top 5 motifs (in terms of counts) occurring in MSAs built from proteins identified in human and mouse PSIBLAST data respectively (Tables-15 and 16).

Table-15: Top 5 motifs found in MSAs built from human and fly proteins in human PSIBLAST data. Numbers in brackets indicate total number of motifs for the given category.

Ran k	Human protein motifs (8831)	Count huma n protei n motifs	Fly protein motifs (15473)	Count fly protei n motifs

1	xxxHTGE K PYxCx xC	48	LRVVRVA K VGRVL RL	51
2	xxxHTGE K PYxCx xC	47	KSIDRQR K LEEALL L	22
3	xxxHTGE K PYxCx EC	46	QSLLDTT K AQVKDI L	22
4	xxxHTGE K PxxCx EC	44	IIDQFHT K ILNDERQ	21
5	xxxHTGE K PYxCx xx	43	PDLLDWR K ARNDR PR	21

Table-16: Top 5 motifs found in MSAs built from mouse and fly proteins in mouse PSIBLAST data. Numbers in brackets indicate total number of motifs for the given category.

Ran k	Mouse protein motifs (266)	Count mouse protei n motifs	Fly protein motifs (1041)	Count Fly protei n motifs
1	AAPAPx E KxPx K K KA	4	xEExx F P K ATDxTFx	17
2	xxxxGLx K xxGxSNF	3	VxxxGx M K FxQxxx x	17
3	EADLVx A K EANx K CP	3	LEAFGN A K TVxND NS	17
4	VSLxAL K K xLAAx	3	xFxEx S A K xxxxNVxx	16

	GY			
5	QAVLLPKKxxxxxx x	3	xxLExQxKELxxxLx	16

While analyzing the tables given above, it is important to note that the count values are actually the total number of 15-mers in a given cluster. For an amino acid to be included in the motif sequence, it should have frequency greater than or equal to 70% of the count value at a given 15-mer position. Failing this criterion, a dummy amino acid x is added at that position in the motif sequence.

All the top 5 15-mer clusters from human proteins (Table-15) have the motif sequence HTGEKPYxCxxC. The MSA from zinc finger proteins contains 5 conserved repeats of the HTGEKPYxCxxC motif. Hence, there are 5 different instances of the same motif (Table-15). In addition to this motif, zinc finger proteins contain another motif sequence CxxCGKxF. However the CxxCGKxF motif occurs with a lower count as compared to HTGEKPYxCxxC motif.

Motif sequence with rank-1 from fly proteins (Table-15) occurs in MSA built from sodium channel proteins. Motifs ranked-2, 3, 4 and 5 occur in MSA built from short stop proteins. Short stop proteins are cross-linkers between F-actin and microtubules. Similar to human proteins, fly proteins also contain zinc finger proteins and HTGEKPYxCxxC motifs albeit with a lower frequency.

Motifs ranked-1, 4 and 5 under mouse column (Table-16) occur in MSAs built from histones. Motif ranked-2 occurs in aldo-keto reductase family 1. Motif ranked-3 occurs in chromobox proteins.

Motifs ranked-1, 2, 3 and 5 in fly columns (Table-16) occur in MSA from myosin heavy chain. Motif rank-4 occurs in Rab proteins.

3.3.5 New predictions made using motifs obtained from protein sequence clusters

A total of 13,922 out of 22,005 proteins of the fruit fly proteome were not picked up by either human or mouse PSIBLAST data. Hence, 15-mers centered on all lysine residues in these 13,922 proteins were extracted. Each of these 15-mers were scanned against a list of 11,227 motifs obtained after combining all fly motifs that had a count equal or greater than 2. The list of 11,227 motifs was created such that if one motif is a subset of another motif, then the larger motif was retained whereas the smaller motif was removed. This was done to ensure that the list of 11,227 motifs was non-repeating and unique in nature. For every query 15-mer the longest motif that matched the query 15-mer was found. This exercise led to identification of 2694 15-mers centered on new lysines from 1131 out of 13,922 proteins. These 2694 15-mers matched 1208 fly motifs. Given below are top 5 fly motifs that matched the most number of 15-mers centered on new lysines (Table-17).

Table-17: Top 5 fly motifs that matched most number of 15-mers centered on new lysines from remaining fly proteome. Numbers in brackets indicate total number of motifs for the given category.

Rank	Fly motifs (1208)	New lysine count (2694)
1	xVxxxxSKSxxxxxx	151

2	xxxxxxPKxxxNxKx	60
3	VxxxDxxKxxxVxxx	55
4	xxxxDxNKDxxxxxx	48
5	xHxxPxVKxxxxxxx	42

Table-18: Summary of proportion of new lysines that conform to consensus motif.

Motif	Proportion of new lysines
[ILV]-K-x-E	121 (5 %)
E-x-K-[ILV]	24 (1 %)
E-x-K-x-E	3 (0 %)
None	2546 (94 %)

Around 6 % of the lysines detected using motifs derived from protein MSAs conform to either forward or inverse orientation of the consensus motif (Table-18).

Table-19: Top 5 proteins and their respective UniProt / UniRef90 IDs for every overlap category.

Ran k	Proteins found by motif matching (UniProt ID : protein name)	Overlap with other fly studies
1	O97159 : Chromodomain-helicase-DNA-binding protein Mi-2 homolog	2
2	P36179 : Serine/threonine-protein phosphatase PP2A 65 kDa regulatory subunit	2

3	Q9NFP5 : SH3 domain-binding glutamic acid-rich protein homolog	2
4	Q9VPH7 : Eukaryotic peptide chain release factor subunit 1	2
5	A0A0B4LH53 : N-ethylmaleimide-sensitive factor 2, isoform B	1

The list of top 5 proteins shown in 2nd column (Table-19) was obtained as follows. First, overlap (with respect to the 4 fly proteomics studies discussed earlier in Venn diagrams) was calculated for each of the 1131 proteins detected from motif matching exercise. Second, this list of 1131 proteins was sorted in descending order by their overlap values. From this sorted protein list, first 5 proteins that contain a lysine found using motif matching were chosen as long as the lysine conformed to either forward or inverse consensus motif.

3.3.6 Analysis of Gene Ontology terms

Given below is the summary of top 5 clusters obtained after analyzing Gene Ontology cellular component, molecular function and biological process terms from human PSIBLAST data (Tables-20 to 22).

Table-20: Top 5 GO C terms found in human and fly proteins in human PSIBLAST data. Numbers in brackets indicate total number of GO C terms for the given protein category.

Rank	GO C terms human	Count : Proportion	GO C terms fly	Count : Proportion
------	------------------	--------------------	----------------	--------------------

	proteins (1206)	%	proteins (902)	%
1	Nucleus	2291 : 11%	Nucleus	2333 : 14%
2	Nucleoplasm	1886 : 9%	Cytoplasm	1531 : 9%
3	Cytosol	1724 : 8%	integral component of membrane	867 : 5%
4	Cytoplasm	1324 : 6%	Cytosol	703 : 4%
5	extracellular exosome	632 : 3%	plasma membrane	545 : 3%

Table-21: Top 5 GO F terms found in human and fly proteins in human PSIBLAST data. Numbers in brackets indicate total number of GO F terms for the given protein category.

Rank	GO F terms human proteins (1991)	Count : Proportion %	GO F terms fly proteins (1510)	Count : Proportion %
1	metal ion binding	1012 : 6%	ATP binding	1196 : 7%
2	DNA binding	821 : 5%	metal ion binding	640 : 4%
3	RNA binding	809 : 5%	zinc ion binding	516 : 3%

4	ATP binding	721 : 4%	RNA binding	493 : 3%
5	DNA-binding transcription factor activity	422 : 3%	DNA binding	480 : 3%

Table-22: Top 5 GO P terms found in human and fly proteins in human PSIBLAST data. Numbers in brackets indicate total number of GO P terms for the given protein category.

Rank	GO P terms human proteins (7037)	Count : Proportion %	GO P terms fly proteins (3855)	Count : Proportion %
1	regulation of transcription, DNA-templated	455 : 1%	regulation of transcription, DNA-templated	393 : 1%
2	positive regulation of transcription by RNA polymerase II	441 : 1%	positive regulation of transcription by RNA polymerase II	333 : 1%

3	negative regulation of transcription by RNA polymerase II	392 : 1%	regulation of transcription by RNA polymerase II	289 : 1%
4	regulation of transcription by RNA polymerase II	283 : 1%	protein phosphorylation	240 : 1%
5	negative regulation of transcription, DNA-templated	268 : 1%	negative regulation of transcription by RNA polymerase II	237 : 1%

As can be seen from the tables above (Tables-20 to 22), majority of the proteins from human PSIBLAST data localize to the nucleus. Most of these proteins bind DNA, RNA or ATP. Regulating transcriptional activity of RNA polymerase II seems to be the common biological process of these proteins.

Given below is the summary of top 5 clusters obtained after analyzing Gene Ontology cellular component, molecular function and biological process terms from mouse PSIBLAST data (Tables-23 to 25).

Table-23: Top 5 GO C terms found in mouse and fly proteins in mouse PSIBLAST data. Numbers in brackets indicate total number of GO C terms for the given protein category.

Rank	GO C terms mouse proteins (491)	Count : Proportion %	GO C terms fly proteins (414)	Count : Proportion %
1	Nucleus	288 : 11%	Nucleus	668 : 17%
2	Nucleoplasm	184 : 7%	Cytoplasm	357 : 9%
3	Cytosol	167 : 6%	Cytosol	187 : 5%
4	Cytoplasm	156 : 6%	plasma membrane	127 : 3%
5	Mitochondrion	65 : 3%	integral component of membrane	95 : 3%

Table-24: Top 5 GO F terms found in mouse and fly proteins in mouse PSIBLAST data. Numbers in brackets indicate total number of GO F terms for the given protein category.

Rank	GO F terms mouse proteins (666)	Count : Proportion %	GO F terms fly proteins (510)	Count : Proportion %
1	identical	99 : 4%	ATP binding	383 : 9%

	protein binding			
2	DNA binding	94 : 4%	RNA polymerase II cis-regulatory region sequence-specific DNA binding	218 : 5%
3	RNA polymerase II cis-regulatory region sequence-specific DNA binding	87 : 4%	DNA-binding transcription factor activity, RNA polymerase II-specific	205 : 5 %
4	ATP binding	77 : 3%	zinc ion binding	205 : 5%
5	metal ion binding	72 : 3%	metal ion binding	137 : 3%

Table-25: Top 5 GO P terms found in mouse and fly proteins in mouse PSIBLAST data. Numbers in brackets indicate total number of GO P terms for the given protein category.

Rank	GO P mouse terms proteins (2219)	Count : Proportion %	GO P terms fly proteins (1690)	Count : Proportion %
1	regulation of transcription by RNA polymerase II	93 : 2%	regulation of transcription by RNA polymerase II	252 : 4%
2	positive regulation of transcription by RNA polymerase II	84 : 2%	regulation of transcription, DNA-templated	160 : 2%
3	negative regulation of transcription by RNA polymerase II	83 : 2%	positive regulation of transcription by RNA polymerase II	86 : 1%
4	negative regulation of transcription, DNA-templated	59 : 1%	negative regulation of transcription by RNA polymerase II	73 : 1%

5	positive regulation of transcription, DNA-templated	49 : 1%	negative regulation of transcription, DNA-templated	60 : 1%
---	---	---------	---	---------

As can be seen above (Tables-23 to 25), GO term analysis of mouse PSIBLAST data reveals that most of the proteins localize to the nucleus.

Most of these proteins bind metal ions, ATP, DNA or RNA. Regulating the transcriptional activity of RNA polymerase II seems to be the primary biological function of these proteins. Thus, taken together from GO term analysis of human and mouse PSIBLAST data, it seems that both categories of proteins have similar biological roles in a cellular environment.

3.3.7 Comparison between predictions from our study, GPS-SUMO and JASSA

The list of SUMOylated lysines predicted in this study was compared with the list obtained from sequence-based SUMOylation site prediction tools namely GPS-SUMO and JASSA. The fly proteins discussed in the present study could come from either fly proteome sequences or the UniRef90 database. Hence, all the 22,005 proteins of the fruit fly proteome and 1087 UniRef90 sequences were submitted as input to GPS-SUMO and JASSA. GPS-SUMO predicted 34,662 SUMOylation sites in 13,216 proteins from the fruit fly proteome and 3518 sites in 766 UniRef90 sequences. JASSA predicted 268,499 lysines in 21,034 proteins from the fruit fly proteome and 26,592 lysines in 1,077 UniRef90 sequences.

Given below is a comparison between lysines predicted in this study and lysines predicted by GPS-SUMO and JASSA (Table-26).

Table-26: Overlap between lysines predicted in this study and predictions made by GPS-SUMO and JASSA.

	Overlap of our study with results from GPS-SUMO	Overlap of our study with results from JASSA	Overlap between results from GPS-SUMO and JASSA
Proteome	2841 (8%)	17,662 (7%)	31,319 (90%)
UniRef90	363 (10%)	2393 (9%)	3181 (90%)

The method presented in this study takes protein homology information into account. However, GPS-SUMO [30,31] and JASSA [16] do not consider homology information while predicting SUMOylation sites. This could be the reason for the low overlap between SUMOylated lysines predicted in this study and those predicted by GPS-SUMO and JASSA. Both GPS-SUMO and JASSA take local sequence environment around lysine residues into account while making predictions and prefer lysines conforming to consensus motifs. Hence, predictions made by both the tools are 90% similar.

3.4 Discussion

The work done in this study uses the concept of sequence homology for annotating SUMOylation sites in the proteome of fruit fly *Drosophila melanogaster* using information derived from human (9256) and mouse (964) proteins. The input information for this study was obtained from mass spectrometry-coupled proteomics experiments. The work resulted in prediction of more than 52,000 SUMOylated lysines present in more than 8600 fly proteins encoded by more than 4600 fly genes. Past fly SUMO proteomics experiments have shown that proteins encoded by 100 of the 4600 fly genes get SUMOylated. Future SUMO proteomics experiments need to validate SUMOylation of proteins encoded by the remaining 3500 fly genes that have been predicted in this study.

Clustering methods discussed in this study provide three kinds of information – sequence motifs centered on SUMOylated lysines, protein family specific motifs and biological functions preferred by target proteins.

First clustering exercise revealed amino acid preferences found in local sequence environment provided by 15-mer s. This analysis confirmed the importance of the ψ -K-x-(E/D) consensus motif. Apart from the consensus motif, the present study also found the existence of hundreds of other motifs in the vicinity of SUMOylated lysines (Tables-11 to 14). Future experiments need to study the biological significance of these motifs because existing scientific literature does not explain the importance of these motifs.

Second clustering exercise revealed protein family specific preferences found with the help of multiple sequence alignments. This analysis showed that members of hundreds of protein families get SUMOylated. Notable examples include zinc finger proteins, sodium channel proteins, histones and chromobox containing proteins. Each protein family has its own set of conserved SUMOylated lysines

and these lysines follow family specific signature motifs. For example, SUMOylated lysines in zinc finger proteins follow a signature motif HTGEKPYxCxxC, sodium channel proteins contain LRVVRVAKVGRVLRL motif and so on (Tables-15 and 16). These family specific motifs in turn helped in predicting SUMOylated lysines in proteins that were missed by PSIBLAST (Tables-17 to 19).

Third clustering exercise helped in understanding cellular and molecular functions of SUMOylation target proteins as revealed by Gene Ontology term analysis. This analysis showed that most of the SUMOylated proteins localize to nucleus, bind DNA / RNA and are involved in regulation of transcriptional activity (Tables-20 to 25).

Future studies could extend this idea to other post-translational modifications such as phosphorylation, ubiquitination, acetylation and others. There are many novel findings of this *in silico* study that need to be experimentally validated in the near future. Such a combination of computational and experimental methods could be used for a better understanding of post-translational modifications in the near future.

Chapter 4: SUMO-ON-THE-FLY web server

4.1 Introduction

Databases are important for storing important biological information such as protein sequences, nucleic acid sequences, atomic structures of biomolecules etc. Examples of such databases include UniProt KB [7] and FlyBase [43] among others. There are also databases of post translational modification sites (such as SUMOylation) determined by low throughput experiments, such as dbPTM [8], PLMD [9], PhosphoSitePlus [10] etc.

Recently, protein sequence alignments were used to identify SUMOylation sites conserved between human and fly proteins as well as mouse and fly proteins respectively. Here, fly refers to the fruit fly *Drosophila melanogaster*. The SUMO-ON-THE-FLY server was designed to make the homology annotations available to fly geneticists around the globe. The database not only predicts known fly SUMOylation sites but also predicts thousands of previously unknown SUMOylation sites.

4.2 Server description

The web server has two separate search forms for the two different kinds of homology data. The database can be searched using UniProt, UniRef90, gene symbol, CG-number or FlyBase identifiers. Search results contain a table wherein the rows refer to mappings between a known SUMOylated lysine in a human / mouse protein and a putative SUMOylated lysine in a fly protein. The result table also contains information related to the sequence alignment between the homologous proteins such as percent identity, percent similarity, alignment length, sequence window centered on the SUMOylation site etc. A screenshot of the server search page is given below (Figure-1). The server has been locally set up at the IP address 10.30.1.175:/sumo-server.com/public_html/, in a Linux environment.

Apache and PHP provide the web interface whereas the database is stored in MySQL.



Figure 1: A screenshot of the search page of the SUMO-ON-THE-FLY server.

4.3 Case study: 14-3-3 epsilon

The database is arranged according to mappings between lysine residues from human / mouse proteins and lysines from fly proteins. For example, when the database is searched for “14-3-3 epsilon”, the database returns information about human and fly homologs, their UniProt identifiers, the position of lysine residues that got aligned etc (Table-1).

Table-1: Results from the database when queried for “14-3-3 epsilon” protein

Fly UniProt ID	Fly lysine	Fly 15-mer	Human UniProt ID	Human lysine	Human 15-mer
P92177	50	NLLSVAYK N VIGARR	P63104	49	NLLSVAYK N VVGARR
P92177	142	FATGSDR K DAAENSL	P63104	139	VAAGDD K KGIVDQSQ
P92177	123	ESKVFY Y KMKGDYHR	P63104	120	ESKVFY L KMKGDYR
P92177	69	IITSIE Q KEENKGAE	P63104	68	VVSSIE Q KTEGAEKK
P92177	215	TLSEES Y KDSTLIMQ	P63104	212	TLSEES Y KDSTLIMQ

GPS-SUMO [14,31] and JASSA [16] are two of the most popular SUMOylation site prediction tools. When the protein sequence of 14-3-3 epsilon from the fruit fly was submitted to the standalone version of GPS-SUMO, it did not predict any of the lysines to be SUMOylated. JASSA predicted many lysines as putative

SUMOylation sites (Table-2). Except lysine 69, none of the lysines predicted by JASSA are predicted by SUMO-ON-THE-FLY database. The 2 lists of predicted lysines differ because unlike SUMO-ON-THE-FLY, JASSA does not take homology information into account while making the predictions. Lysines at positions – 12, 73, 78, 83, 118, 125 and 250 were not predicted by SUMO-ON-THE-FLY because these lysines were not detected as SUMOylation sites by the mass spectrometry-coupled proteomics experiments used in chapter 3 of this thesis.

Table-2: SUMOylation sites in 14-3-3 epsilon protein as predicted by JASSA and the 21-mers centered on these lysines as returned by JASSA

Lysines predicted by JASSA	21-mers centered on those lysines
12	TERENNVYKAKLAEQAERYDE
69	SWRIITSIEQKEENKGAEKLE
73	ITSIEQKEENKGAEKLEMIK
78	QKEENKGAEKLEMIKTYRGQ
83	KGAEKLEMIKTYRGQVEKEL
118	LIPCATSGESKVFYYKMKGDY
125	GESKVFYYKMKGDYHRYLAEF
250	VDPNAGDGEPKEQIQDVEDQD

4.4 Acknowledgements

The user interface of the server was designed by Neelesh Soni, whereas the MySQL databases were generated by Yogendra Ramtirtha.

Chapter 5: 3-D structure based prediction of SUMOylation sites

5.1 Introduction

Lysine residues in eukaryotic proteins are known to undergo many post-translational modifications. Examples include ubiquitination, acetylation, methylation, SUMOylation etc. SUMOylation involves formation of a covalent bond between side chain amino group of lysine residues in target / substrate proteins and C-terminus of a protein called SUMO (Small Ubiquitin-related MOdifier). Candidate lysines from target proteins are selected by SUMO E2 conjugating enzyme (ubc9) independently *in vitro* or with the help of SUMO E3 ligases *in vivo*. Mutation of lysine to an arginine disrupts the modification. Disruption of SUMOylation has been linked to neuro-degenerative diseases and cancer.

Experimental determination of SUMOylated lysines is cumbersome. Hence, computational prediction of SUMOylated lysines could be useful. All the currently available SUMOylation site prediction tools make use of protein sequences. GPS-SUMO [15,44] and JASSA [16] are two such popular sequence-based SUMOylation site prediction tools. These sequence-based SUMOylation site prediction tools preferentially look for the consensus motif ψ -K-x-(E/D), where ψ – I/L/V and K – SUMOylated lysine. Recent mass spectrometry-coupled proteomics experiments conducted on human cell lines have shown that around 50% of all SUMOylated lysines conform to consensus motif [11,13]. Thus, protein sequence information alone is insufficient to predict all SUMOylated lysines. Information about protein three dimensional (3-D) structures could be useful to understand how ubc9 discriminates between SUMOylated and non-SUMOylated lysines. To the best of our knowledge, none of the currently available SUMOylation site prediction tools make direct use of protein 3-D information. Hence, this research article describes proof-of-concept study of a novel method that takes protein 3-D information into account while predicting SUMOylated lysines.

Experimental techniques such as x-ray crystallography, nuclear magnetic resonance and electron microscopy are used to solve protein structures. Protein 3-D structures are deposited in a data archive called Protein Data Bank (PDB) [45] under different accession identifiers. Structural information about known ubc9-target protein complexes is very important for designing a robust prediction method. Our current understanding of ubc9-target complexes is limited because the PDB contains information about only one ubc9-target complex where the target protein is RanGAP1 (PDB ID : 1Z5S) [46]. Given below is the image of the enzyme-target complex generated using UCSF Chimera version 1.13.1 [47] (Figure-1). Ubc9 active site has 2 important residues C93 and D127 that catalyze the formation of covalent bond between lysine side chain and SUMO C-terminal tail. Lysine binding site in ubc9 is so narrow that replacing lysine in the target protein with an arginine disrupts the formation of covalent bond.

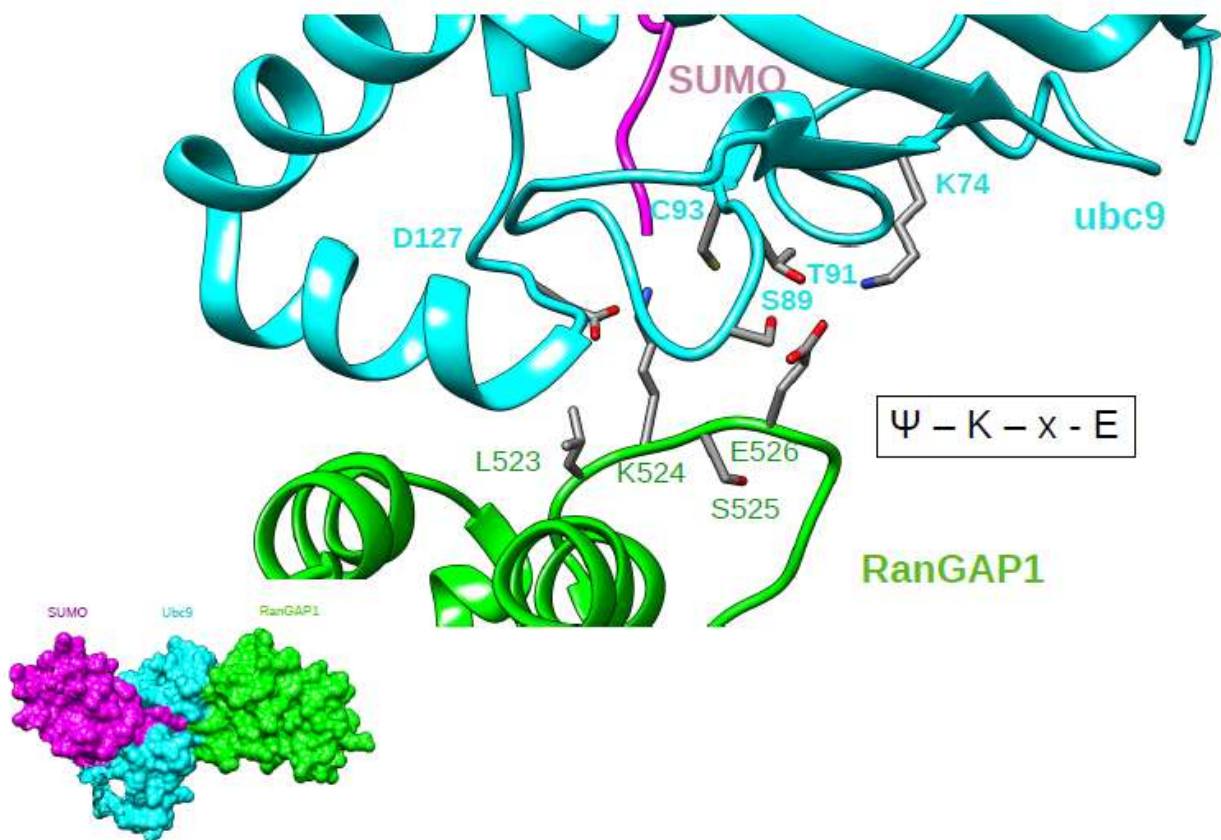


Figure 1 : Top : Ribbon representation of a complex between SUMO (magenta), ubc9 (cyan) and RanGAP1 (green). All interacting residues are shown in ball and stick models. inset : surface representation of the same complex and same coloring scheme. SUMOylated lysine from RanGAP1 conforms to the consensus motif.

Scarcity of structural information about ubc9-target complexes is the major limiting factor for the present study. Hence, this study has been divided into three broad steps. First step involves creating a dataset of protein 3-D structures for SUMOylated proteins identified by recent experiments. All the proteins considered in this study are encoded of human origin. Second step involves docking target protein structure onto ubc9 structure such that there is a lysine near the active site of ubc9. Different conformational poses of ubc9-target complex are sampled and the pose with maximum number of inter-protein atomic contacts at a distance of 4Å between the two proteins is chosen. Care was taken to make sure that the chosen pose did not have any atomic clashes between the main chain atoms of both the proteins. The sampling method was applied to every lysine from all the structures of the dataset and optimal pose of the ubc9-target complex obtained for every lysine was chosen. The third step involves developing a scoring method that can discriminate between ubc9-target complexes of SUMOylated and non-SUMOylated lysine. Performance of the scoring method will also be assessed/

5.2 Materials and methods

5.2.1 Generation of a dataset of SUMOylated protein structures

The list of SUMOylated proteins for this study was obtained from a recent mass-spectrometry based proteomics experiments conducted on human cell lines [13]. This list consists of 9330 proteins containing 49850 SUMOylated lysines in total. The mappings between human SUMOylated proteins and their respective Protein Data Bank (PDB) [45] identifiers were obtained with the help of SIFTS database [48,49]. Around 2331 of the 9330 SUMOylated proteins have at least one structure in the PDB containing at least one SUMOylated lysine. In order to remove redundancy in these proteins, h-CD-HIT server (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=h-cd-hit) [35–38] was used with 3 hierarchical identity cutoffs 90%, 60% and 30% respectively. The results obtained at 30% redundancy from h-CD-HIT server contained a list of 1841 structures corresponding to 1841 SUMOylated proteins. The details of dataset used in this study are given below (Table-1). Some protein structures contained unnatural amino acids such as seleno-methionine, phospho-serine, phospho-threonine, phospho-tyrosine etc. These unnatural amino acids would have created errors

during the sampling method discussed later. Hence, all the unnatural amino acids were converted to their nearest analogues from the 20 standard amino acids such as methionine, serine, threonine, tyrosine etc. Protein structures often have missing atoms because some parts of the structure have poor resolution. All such missing atoms were fixed using `complete_pdb()` function in MODELLER version 9.17 [39]. In some cases, residue numbering differs between UniProt sequence of a protein and the sequence of its corresponding PDB structure. In order to obtain correct residue positions, pairwise alignments were built between UniProt and PDB sequences of each protein. All pairwise alignments were built using SALIGN from MODELLER version 9.17. There are around 7432 SUMOylated lysines in the 1841 structures used in this study. All the remaining 27874 lysines in these 1841 protein structures (except the 7432 lysines) were treated as non-SUMOylated lysines.

Table-1: Overview of dataset used in this study

Description	Numbers
Total number of target proteins	1841
Total number of SUMOylated lysines	7432
Lysines conforming to either K-x-(E/D) or (E/D)-x-K motif	2556
Lysines not conforming to consensus motif	4876

5.2.2 Computational tools used in this study

All the steps in this work including dataset compilation, sampling method and scoring analysis were implemented in Python version 2.7.5. Mathematical calculations were carried out using Numeric Python (NumPy) version 1.7.1 [40]. Graphs were plotted using ggplot2 library [50] in R version 3.4.4 [42].

5.2.3 Sampling method to dock target proteins onto ubc9

It is important to understand the structure of a lysine residue before discussing the sampling method. A lysine residue has 4 main chain atoms N, CA, C and O as well as 5 side chain atoms CB, CG, CD, CE and NZ (Figure-2). The lysine main chain

has 2 torsion angles phi and psi whereas its side chain has 4 torsion angles chi1, chi2, chi3 and chi4 (Figure-2). Angles between 4 atoms connected by 3 consecutive bonds are known as torsion angles.

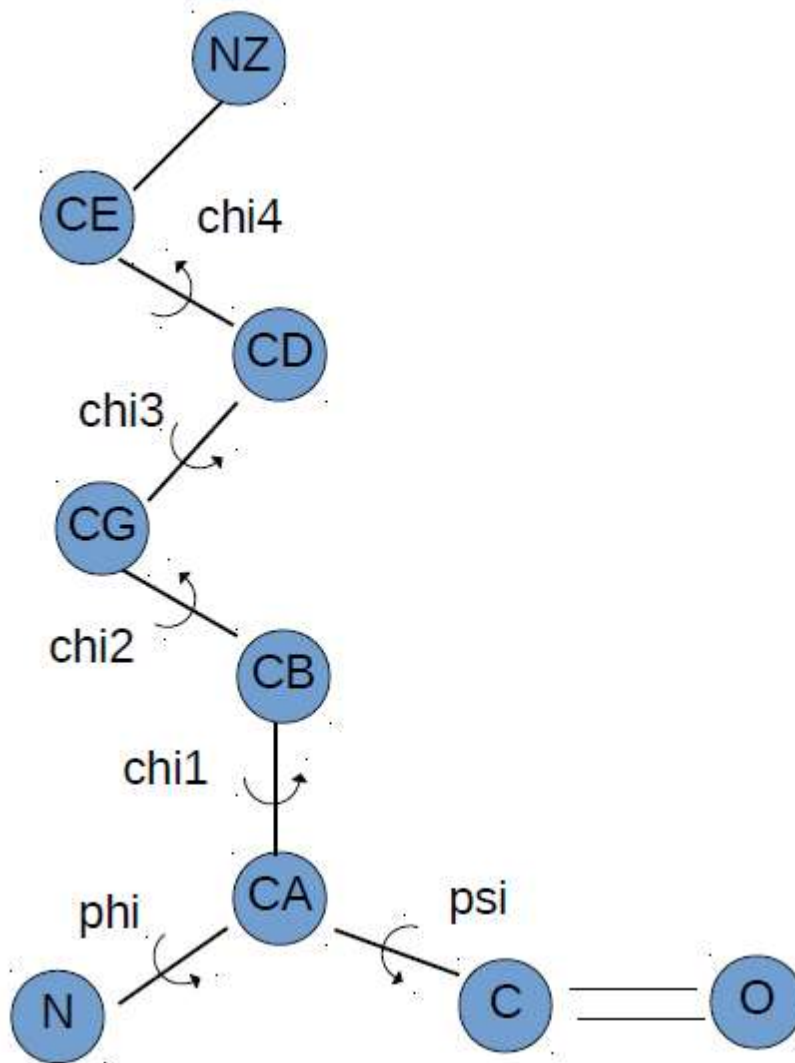


Figure 2: Two dimensional structure of a lysine residue

Table-2: Overview of atoms involved in different torsion angles

Torsion angle	Atoms involved	Bond of interest
Phi	C_{prev} -N-CA-C	N-CA
Psi	N-CA-C- N_{next}	CA-C
chi1	N-CA-CB-CG	CA-CB
chi2	CA-CB-CG-CD	CB-CG
chi3	CB-CG-CD-CE	CG-CD

chi4	CG-CD-CE-NZ	CD-CE
------	-------------	-------

Details of atoms involved in different torsion angles are given above (Table-2). For example, chi4 angle measures rotation of atoms around CD-CE bond and involves atoms CG, CD, CE and NZ. C_{prev} and N_{next} imply main chain C and N atoms of previous and next residues in the protein sequence respectively.

Torsion angle calculations are important for the sampling method. These calculations depend on unit vector calculations. For a vector $v = (x, y, z)$, unit vectors are calculated in 2 steps. First, modulus of v is calculated by taking the square root of a dot product of v with itself (Equation-1). Second, v is divided by the modulus to obtain a unit vector of v (Equation-2). In Python, dot product is calculated using the function `numpy.dot()` and square root is calculated using `numpy.sqrt()` function.

Equation 1

$$\text{modulus of } v = \sqrt{v \cdot v}$$

Equation 2

$$\text{unit vector of } v = \frac{v}{\text{modulus of } v}$$

In order to understand torsion angle calculations, let us consider the chi4 angle. The chi4 angle measures rotation of atoms around CD-CE bond and it measures angle between two planes. The first plane is formed by CG, CD and CE atoms whereas the second plane is formed by CD, CE and NZ atoms. In case the four atoms have the Cartesian coordinates – $CG = (x1, y1, z1)$, $CD = (x2, y2, z2)$, $CE = (x3, y3, z3)$ and $NZ = (x4, y4, z4)$, then the torsion angle calculations go as follows:

First, we calculate vectors: $b1 = (x1 - x2, y1 - y2, z1 - z2)$, $b2 = (x3 - x2, y3 - y2, z3 - z2)$ and $b3 = (x4 - x3, y4 - y3, z4 - z3)$. Second, $v1$, $v2$ and $v3$ are unit vectors along $b1$, $b2$ and $b3$ respectively. Third, vectors $u1$ and $u3$ are calculated (Equations-3 and 4).

Equation 3

$$u1 = b1 - (b1.v2)v2$$

Equation 4

$$u3 = b3 - (b3.v2)v2$$

Equation 5

$$cs = u1.u3$$

Equation 6

$$sn = u1 \times u3$$

Equation 7

$$sn = \text{Fobrenius norm}(u1 \times u3)$$

Equation 8

$$ang = \text{atan2}(sn, cs)$$

Equation 9

$$sign = (u1 \times u3).v2$$

In all the above equations note the difference between dot product (.) and cross product (x). Fobrenius norm for a matrix is obtained by taking the square root of the sum of squares of all elements of the matrix. In Python, Fobrenius norm is calculated by using the function `numpy.linalg.norm()` and `atan2` is calculated by using the function `numpy.arctan2`. The term “ang” represents the value of the torsion angle in radians. If the sign term is negative, then ang is also negative and hence must be multiplied by -1. The torsion angle “ang” ranges between $-\pi$ to π in radians or -180° to 180° . In order to convert “ang” from radians to degrees, multiply “ang” by a factor of $180/\pi$. Conversely, in order to convert an angle from degrees to radians, it should be multiplied by a factor of $\pi/180$.

Given above are mathematical equations necessary for torsion angle calculations (Equations-1 to 9). Now we will be discussing the mathematical equations necessary for calculating Cartesian coordinates of a rigid body after rotation. For example, let us consider 3 points – $P1 = (x1, y1, z1)$, $P2 = (x2, y2, z2)$ and $Q = (x3, y3, z3)$. Axis of rotation passes through points P1 and P2, and we want to calculate

Cartesian coordinates of Q after rotation by an angle θ in radians around the axis of rotation. Vector $b1 = (x2 - x1, y2 - y1, z2 - z1)$ and $u = (x0, y0, z0)$ is a unit vector along the direction of vector $b1$.

Equation 10

$$t1 = x0 * x0 * (1 - \cos\theta) + \cos\theta$$

Equation 11

$$t2 = y0 * x0 * (1 - \cos\theta) - z0 * \sin\theta$$

Equation 12

$$t3 = z0 * x0 * (1 - \cos\theta) + y0 * \sin\theta$$

Equation 13

$$t4 = x0 * y0 * (1 - \cos\theta) + z0 * \sin\theta$$

Equation 14

$$t5 = y0 * y0 * (1 - \cos\theta) + \cos\theta$$

Equation 15

$$t6 = z0 * y0 * (1 - \cos\theta) - x0 * \sin\theta$$

Equation 16

$$t7 = x0 * z0 * (1 - \cos\theta) - y0 * \sin\theta$$

Equation 17

$$t8 = y0 * z0 * (1 - \cos\theta) + x0 * \sin\theta$$

Equation 18

$$t9 = z0 * z0 * (1 - \cos\theta) + \cos\theta$$

Now, we calculate vector $b2 = (x4, y4, z4)$, where $x4 = x3 - x1$, $y4 = y3 - y1$ and $z4 = z3 - z1$. Let us say that $(x5, y5, z5)$ are Cartesian coordinates of point Q after the rotational motion. The values of $x5$, $y5$ and $z5$ can be calculated as follows.

Equation 39

$$x5 = (t1 * x4 + t2 * y4 + t3 * z4) + x1$$

Equation 20

$$y5 = (t4 * x4 + t5 * y4 + t6 * z4) + y1$$

Equation 21

$$z5 = (t7 * x4 + t8 * y4 + t9 * z4) + z1$$

Equations 19 to 21 can also be summarized as matrix multiplication given below.

$$\begin{matrix} x5 & t1 & t2 & t3 & x4 & x1 \\ y5 & = & t4 & t5 & t6 & * & y4 & + & y1 \\ z5 & & t7 & t8 & t9 & & z4 & & z1 \end{matrix}$$

In Python, the values of $\sin\theta$ and $\cos\theta$ are calculated using the functions `numpy.sin()` and `numpy.cos()`. And, the value of π is obtained using the function `numpy.pi()`. Thus mathematical equations given above help us calculate the Cartesian coordinates of a target protein after spinning around an axis of rotation (Equations-10 to 21).

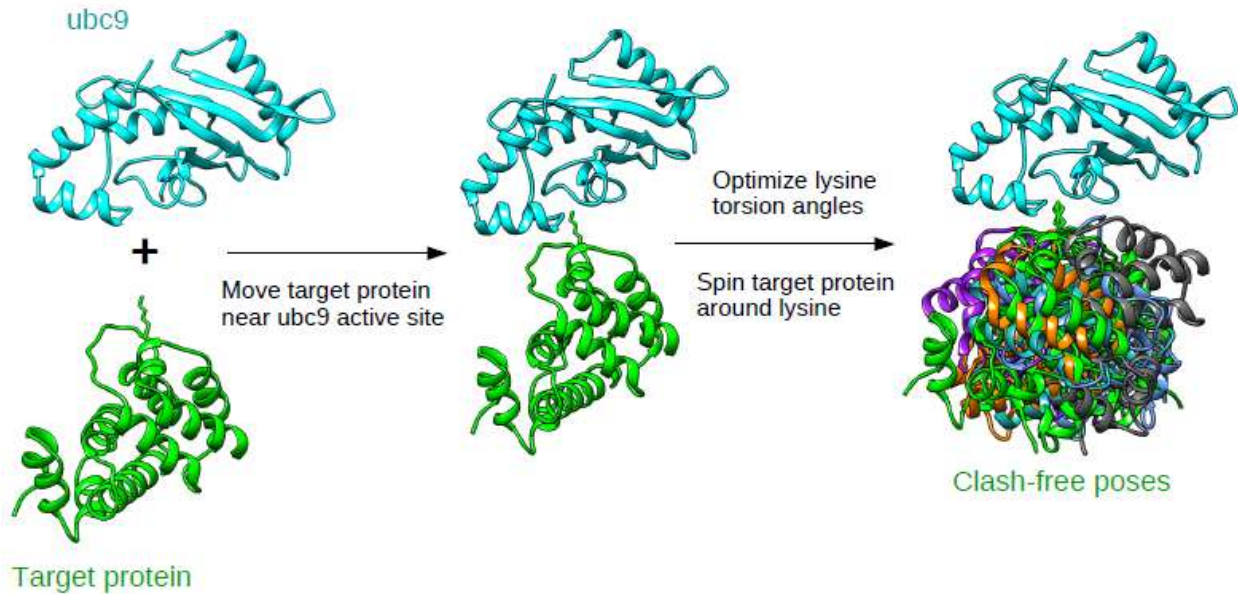


Figure 3: Schematic overview of the steps involved in sampling method. **Left panel** – the method begins with monomeric unbound structures of ubc9 and target protein. **Centre panel** – Lysine residue from target protein is introduced into the active site of ubc9 using a technique called rigid transformation. **Right panel** – Keeping ubc9 fixed and moving the target protein, different clash-free conformational poses are sampled between the two proteins. In the above figure, all clash-free poses are represented as ribbons colored using different colors.

The objective of the sampling method is to dock target proteins onto *ubc9* such that the lysine of interest from the target is near the active site residues of the enzyme. After docking the target onto *ubc9*, the method samples favorable conformational poses between the target and the enzyme. In the present method, *ubc9* and target proteins are treated as rigid bodies. During the sampling process, *ubc9* remains fixed whereas the target protein undergoes motions such as translations and rotations. The sampling method is carried out independently for every lysine from each of the 1841 target proteins of the dataset.

Translation is a motion wherein every atom in a rigid body is displaced by the same distance through 3-D space. Rotation is a motion wherein every atom in a rigid body spins around an axis of rotation. Schematic depiction of steps involved in the sampling method is given above (Figure-3). Each step of the sampling method is elaborated below.

*5.2.3.1 Move target protein near *ubc9* active site*

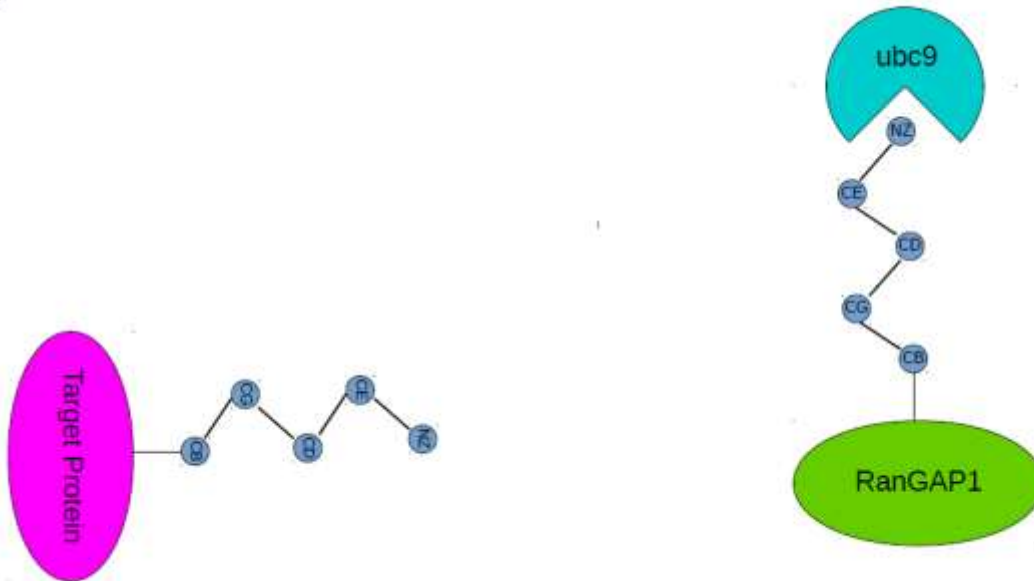
The first step of the sampling method aims at bringing the target protein in the vicinity of *ubc9* such that the lysine of interest from the target is in the active site of the enzyme. This is achieved with the help of a technique called rigid body transformation in 3-D space, also known as 3-D least squares fit. Given below is a pictorial overview of steps involved in rigid body transformation in 3-D space (Figure 3).

A detailed explanation of rigid body transformation can be found here (http://nghiaho.com/?page_id=671) [51]. The structure of protein RanGAP1 bound to *ubc9* (PDB ID: 1Z5S and Figure 1) was used as a reference for the transformation. CD, CE and NZ atoms from lysine-524 in RanGAP1 and lysine of interest in target protein are important for the transformation process. The transformation process minimizes the root mean squared deviation between both the atom sets.

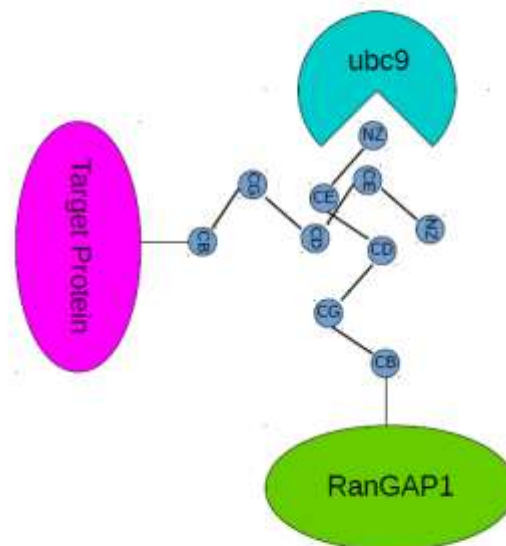
Before the beginning of the transformation, the target and RanGAP1 proteins could have any arbitrary location in 3-D space (Figure 3A). Let us say, $CD1 = (x1, y1, z1)$, $CE1 = (x2, y2, z2)$ and $NZ1 = (x3, y3, z3)$ are side chain atoms of lysine of interest from target protein. And, $CD2 = (x4, y4, z4)$, $CE2 = (x5, y5, z5)$ and $NZ2 =$

(x6, y6, z6) are side chain atoms of lysine 524 from RanGAP1. The first step of the transformation process is to translate the target protein such that the center of masses of both the above mentioned atom sets superimpose (transition of target protein from Figure 3A to 3B).

A



B



c

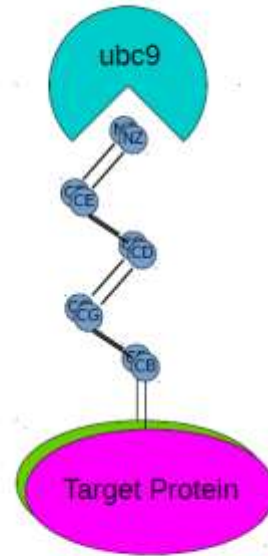


Figure 4: Schematic representation of steps involved in rigid body transformation of target protein in 3-D space.

Given below are mathematical equations important for the transformation. Here, $cm1$ and $cm2$ are center of masses of both the atom sets described above. Average Cartesian coordinates are $xa = \frac{x1 + x2 + x3}{3}$, $ya = \frac{y1 + y2 + y3}{3}$, $za = \frac{z1 + z2 + z3}{3}$ and $xb = \frac{x4 + x5 + x6}{3}$, $yb = \frac{y4 + y5 + y6}{3}$, $zb = \frac{z4 + z5 + z6}{3}$.

Equation 22

$$cm1 = (xa, ya, za)$$

Equation 23

$$cm2 = (xb, yb, zb)$$

Equation 24

$$setA = \begin{matrix} x1 - xa & y1 - ya & z1 - za \\ x2 - xa & y2 - ya & z2 - za \\ x3 - xa & y3 - ya & z3 - za \end{matrix}$$

Equation 25

$$\text{set}B = \begin{matrix} x4 - xb & y4 - yb & z4 - zb \\ x5 - xb & y5 - yb & z5 - zb \\ x6 - xb & y6 - yb & z6 - zb \end{matrix}$$

Equation 26

$$H = \text{transpose}(\text{set}A) * \text{set}B$$

Equation 27

$$U, S, Vh = \text{SVD}(H)$$

Equation 28

$$\text{Rotmat} = \text{transpose}(U) * \text{transpose}(Vh)$$

Equation 29

$$\text{Transmat} = -\text{Rotmat} * \text{transpose}(\text{cm}1) + \text{transpose}(\text{cm}2)$$

All the equations given above (Equations 22 to 29) involve matrix multiplications. “Rotmat” and “Transmat” are rotation and translation matrices that help us in calculating the Cartesian coordinates of a target protein after rigid body transformation. SVD stands for singular value decomposition (implemented using Python function `numpy.linalg.svd()`). After the transformation is completed, the Cartesian coordinates of RanGAP1 were deleted leaving behind the ubc9-target complex.

Rigid body transformations help in docking the target protein onto ubc9 (Figure 3A, 3B and 3C). However, the structure of the enzyme-target complex generated from transformation may not be the energetically favorable pose for the two proteins to interact. Hence, the enzyme-target complex was subjected to further conformational sampling, the details of which are given below.

5.2.3.2 Optimize lysine torsion angles

The chi2, chi3 and chi4 torsion angles of lysine of interest from the ubc9-target complex generated above were adjusted to 172.8°, 173.8° and -175.3° respectively. This was done because the tunnel in ubc9 through which lysine accesses the active

site is very narrow and accommodates lysine only in its stretched or extended conformation (Figure-1). Mutagenesis studies have shown that if lysine is substituted by arginine, then the arginine cannot enter the tunnel owing to its branched side chain. In order to adjust torsion angles to their appropriate values, all the atoms of target protein (except lysine side chain atoms) are rotated around an axis of rotation defined by bond of interest. For example, while adjusting the chi4 angle, all the atoms of target protein are rotated around CD-CE bond (Table-3) except CD, CE and NZ atoms (Table-3) of lysine side chain. Similar procedure is applied to chi1, chi2 and chi3 angles.

Table-3: Lysine side chain torsion angles, bond of interest and atoms kept fixed during angle adjustment

Torsion angle	Bond of interest	Atoms kept fixed during adjustment
chi1	CA-CB	CA, CB, CG, CD, CE, NZ
chi2	CB-CG	CB, CG, CD, CE, NZ
chi3	CG-CD	CG, CD, CE, NZ
chi4	CD-CE	CD, CE, NZ

The list of possible values of chi1 angle was obtained from 2010 back-bone dependent rotamer library [52]. First main chain torsion angles phi and psi are calculated for the lysine of interest. Second, the rotamer library is searched for all possible lysine conformations having the given phi and psi angles as well as having average chi2, chi3 and chi4 angle values within $180 \pm 10^\circ$. This list of chi1 angle values usually consists of 3 approximate conformations: -60° (+gauche), 60° (-gauche) and 180° (trans) respectively. The lysine of interest from the target protein is sampled in all the 3 conformations.

5.2.3.3 Spin target protein around lysine

For every chi1 angle value, the target protein is subjected to an additional rotation. The axis of rotation for this motion is defined by CB and NZ atoms of lysine of interest. The target protein is spun in steps of 10° . Thus, for every chi1 angle value,

there are $360^\circ / 10^\circ = 36$ different conformational poses between ubc9 and target protein. At the end of the sampling process, a total of $3 * 36 = 108$ conformational poses are sampled between ubc9 and target protein for a given lysine of interest.

Out of all the conformational poses sampled between ubc9 and target protein for a given lysine of interest, only those poses are retained that have no clashes between main chain atoms of ubc9 and target protein. Here, N, CA, C, O and CB atoms of both the proteins are considered as main chain atoms. The distance between ubc9 main chain atom A = (x1, y1, z1) and target protein main chain atom B = (x2, y2, 2z) is calculated as follows.

Equation 30

$$distance = \sqrt{(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2}$$

The Lennard-Jones potential (LJ) between atoms A and B can be calculated as follows.

Equation 31

$$LJ\ potential = \epsilon_{A,B} \left(\left(\frac{R_{min,A,B}}{d_{A,B}} \right)^{12} - 2 \left(\frac{R_{min,A,B}}{d_{A,B}} \right)^6 \right)$$

Here, $\epsilon_{A,B}$ and $R_{min,A,B}$ are terms specific to atoms A and B that were obtained from topology parameters of AMBER 99 force field [53]. The $R_{min,A,B}$ term is the sum of van der Waals radii of atoms A and B. When LJ potential between atoms A and B is equal to zero, Equation 31 reduces to $d_{A,B} = 0.89 * R_{min,A,B}$. Thus, atoms A and B are considered to be clashing if the distance between them is less than 0.89 times the sum of their van der Waals radii. At the end of sampling method, only those poses are retained that do not have clashes between main chain atoms of ubc9 and target protein. In case, a lysine of interest has more than one clash free poses, then the pose having maximum number of atomic contacts within a distance of 4\AA between ubc9 and target protein was chosen. This was done to ensure that one representative pose was chosen for every lysine of interest.

5.2.4 Discriminating poses based on residue contacts

The aim of this exercise is to find a combination of residue contacts that occur more in ubc9-target poses of SUMOylated lysines than in poses of non-SUMOylated lysines. This was achieved with the help of a modified version of Apriori algorithm [34]. Residue *i* from ubc9 and residue *j* from a target protein was considered to be in contact if any atom from *i* is within a distance of 4Å from any atom of residue *j*. Residue contact information from all the ubc9-target poses is encoded as “res-pairs”. An example of res-pair encoding for residue contacts from ubc9-RanGAP1 complex (Figure-1) involving glutamate residues of consensus motif from RanGAP1 is given below (Table-4). Residue contacts involving lysine of interest are ignored from all ubc9-target complexes because they do not provide any new information.

Table-4: Example of res-pair encodings for residue contacts between ubc9 and glutamate residue of consensus motif from RanGAP1

Ubc9 residue number : ubc9 residue type – RanGAP1 residue number : RanGAP1 residue type	Res-pair encoding
89 : SER – 526 : GLU	89 – GLU
91 : THR – 526 : GLU	91 – GLU
74 : LYS – 526 : GLU	74 – GLU

The Apriori algorithm is commonly used for finding patterns in customer transaction data in the retail industry. Apriori algorithm clusters different items bought by customers into sets according to their support (also known as probability). For the present exercise, a res-pair could be considered as an item and res-pairs were clustered into different res-pair sets on the basis of their occurrences. The definition of support was modified as follows (Equation 32). The modified version of support was referred to as enrichment.

Equation 32

$$\text{enrichment} = \text{observed probability} - \text{expected probability}$$

Here, observed probability refers to normalized frequency of a res-pair (or a set of res-pairs) in ubc9-target poses derived from SUMOylated lysines. Expected

probability refers to normalized frequency of res-pair (or set of res-pairs) in ubc9-target poses derived from non-SUMOylated lysines. Normalization was done with respect to total number of ubc9-target poses for the given lysine category – SUMOylated or non-SUMOylated.

The Apriori algorithm was applied to 360 res-pairs that occur in ubc9-target poses of both SUMOylated and non-SUMOylated lysines. These 360 res-pairs also had absolute frequency of 40 or higher in poses of SUMOylated lysines. The Apriori algorithm can be summarized in the steps given below –

- i. Calculate enrichment values for each of the 360 res-pairs. Each res-pair could be thought of as a res-pair set of size equal to 1. All res-pairs having their enrichments greater than or equal to -1.0 were selected. (There were no res-pairs having enrichments greater than or equal to 0.0).
- ii. New res-pair sets were generated by extending res-pair sets from previous step by another res-pair, such that the newly added res-pair was a member of a res-pair set from previous step. All possible res-pair sets of size equal to 2 were generated.
- iii. Enrichment values for all newly generated res-pair sets were calculated. All res-pair sets having enrichments greater than or equal to 0.0 were selected.
- iv. This process of generation and selection of new res-pair sets having size greater by 1 than previous step is continued till no new res-pair sets can be created. At this step, the algorithm ends.

5.2.5 Statistical parameters to assess predictions

Predictions were assessed using statistical parameters such as sensitivity, specificity, accuracy, F1 score and MCC (Matthew's Correlation Coefficient) (Equations-33 to 38). Here, TP = number of true positives, TN = number of true negatives, FP = number of false positives and FN = number of false negatives.

Equation 33

$$sensitivity = \frac{TP}{TP + FN}$$

Equation 34

$$specificity = \frac{TN}{TN + FP}$$

Equation 35

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 36

$$precision = \frac{TP}{TP + FP}$$

Equation 37

$$F1\ score = 2 * \frac{precision * sensitivity}{precision + sensitivity}$$

Equation 38

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

5.3 Results

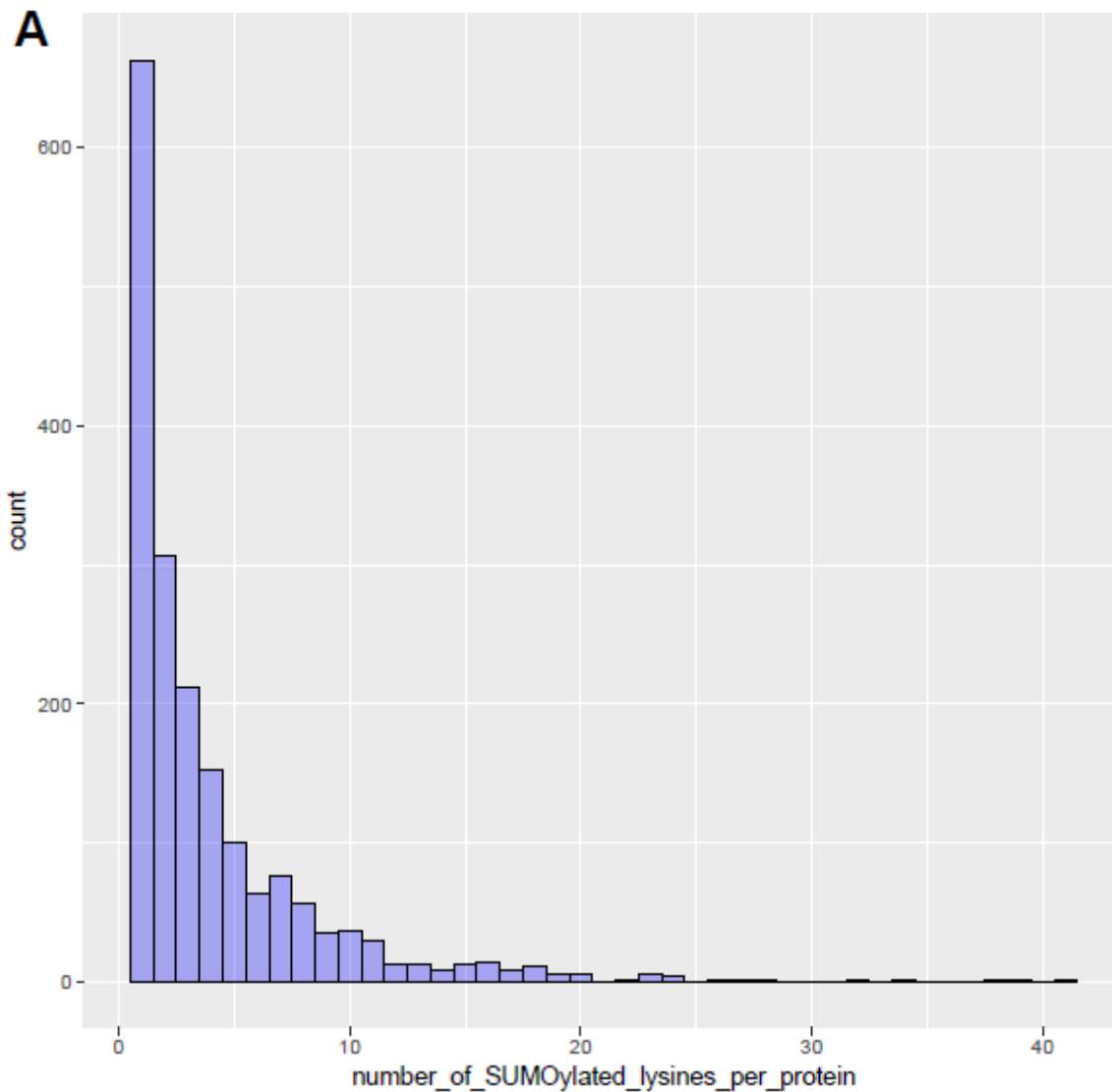
5.3.1 Analysis of the target protein structures in the dataset

Majority (1349) of the protein structures used in this study were solved using x-ray crystallography (Table-5). However, a few of the structures were also solved using NMR (178), electron microscopy (313) and solution scattering (1).

Table-5: Experimental sources of the protein structures used in this study

Experiment type	Number of structures
X-ray crystallography	1349
NMR	178
Electron microscopy	313
Solution scattering	1

The protein structures used in this study vary in their sequence lengths (Figure 5B) as well as the number of SUMOylated lysines (Figure 5A) present in them. Most of the proteins used in this study contain 5 or less SUMOylated lysines. However, there are a handful of proteins that contain as many as 20 or more modified lysines. Similarly, majority of protein structures used in this study have sizes less than 1000 amino acids. There are a handful of structures that have sizes equal to or greater than 2000 amino acids. Thus, SUMOylation occurs in proteins of varying sizes and varying number of lysines in these proteins.



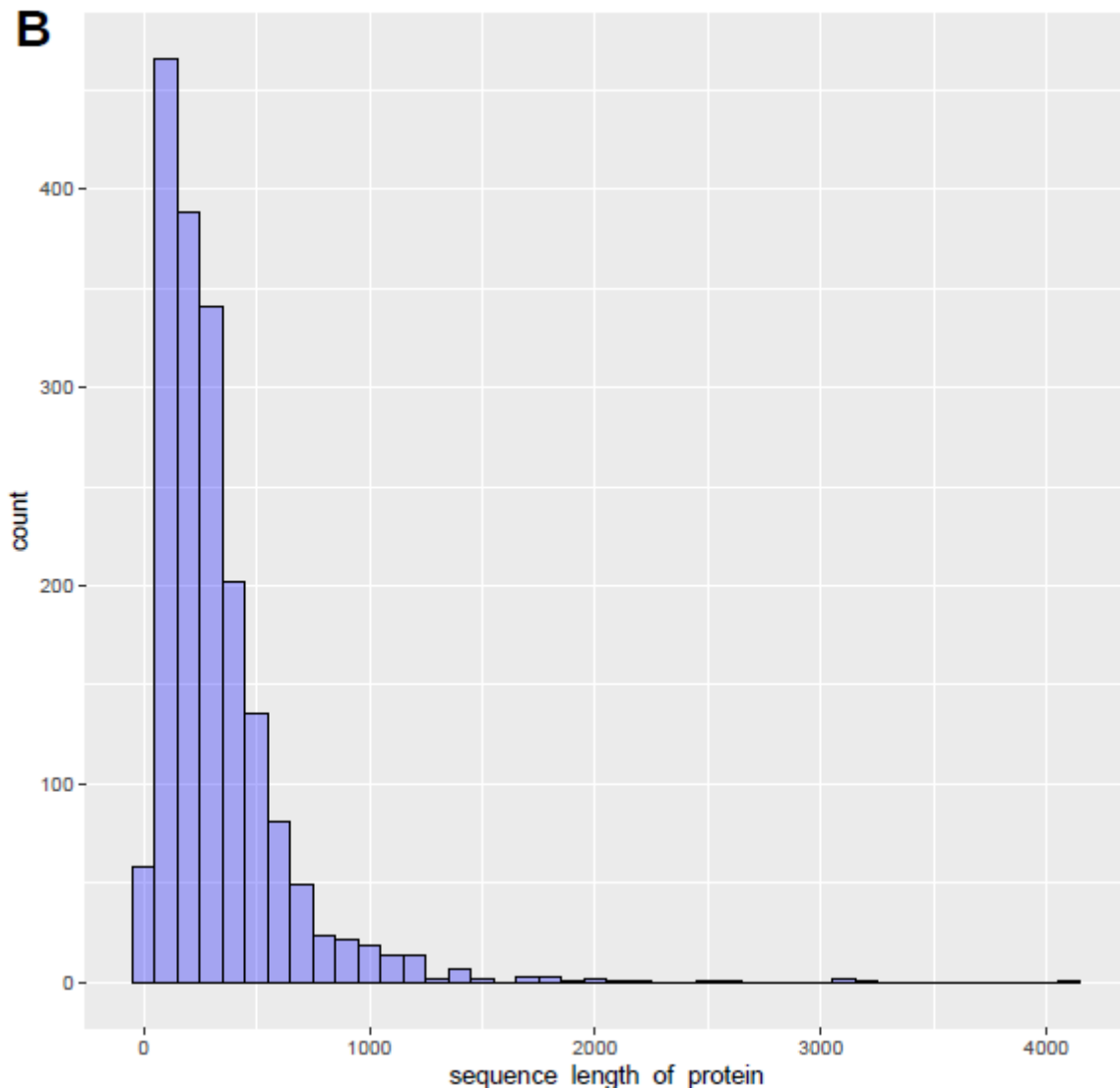


Figure 5: Histograms of **A:** number of SUMOylated lysines per target protein structure and **B:** sequence length of every target protein. Here count refers to frequency or number of proteins.

Table-6: Proportion of SUMOylated lysines that conform to consensus motif

Description	Numbers
Lysines conforming to K-x-(E/D) motif	1174
Lysines conforming to (E/D)-x-K motif	1175
Lysines conforming to (E/D)-x-K-x-(E/D) motif	207
Lysines not conforming to	4876

consensus motif	
-----------------	--

There are around 7618 lysines in the 1841 protein structures that conform to either the forward or reverse version of the K-x-(E/D) consensus motif. However, only 2556 or 33.5% of these lysines tend to get SUMOylated. Thus, a consensus motif alone is insufficient to guarantee SUMOylation of a lysine residue. On the other hand, all the SUMOylated lysines do not necessarily conform to the consensus motif. Majority of SUMOylated lysines analyzed in this study (4876) do not conform to the consensus motif. Thus, a sequence motif alone is not sufficient to predict all the SUMOylated lysines. Hence, the present study uses information from interactions between 3-D structures of unc9 and target proteins to predict SUMOylation sites.

Table-7: Representatives of all the CATH superfamilies are included in this study

CATH superfamily	Count
Mainly Alpha	478
Mainly Beta	338
Alpha Beta	947
Few Secondary Structures	13
Special	20

The CATH database [54] classifies protein structures into different superfamilies (folds). Protein structures of the dataset used in this study were mapped to their respective CATH superfamilies (Table-7). There are 5 CATH superfamilies and members of all of these superfamilies are included in the dataset used in this study. The superfamily Alpha Beta has maximum representation as compared to other superfamilies.

Table-8: Top 10 most abundant cellular component terms

Rank	Gene Ontology cellular component terms	Count (Proportion %)
1	Nucleus	419 (8.5)
2	Nucleoplasm	411 (8.3)
3	Cytosol	393 (7.94)
4	Cytoplasm	325 (6.57)

5	Extracellular exosome	179 (3.62)
6	Membrane	127 (2.57)
7	Plasma membrane	98 (1.98)
8	Nucleolus	80 (1.62)
9	Mitochondrion	70 (1.41)
10	Protein-containing complex	68 (1.37)

Table-9: Top 10 most abundant molecular function terms

Rank	Gene Ontology molecular function terms	Count (Proportion %)
1	RNA binding	151 (3.88)
2	Identical protein binding	148 (3.8)
3	Metal ion binding	102 (2.62)
4	ATP binding	101 (2.6)
5	DNA binding	91 (2.34)
6	Zinc ion binding	66 (1.7)
7	Chromatin binding	62 (1.59)
8	Protein homodimerization activity	56 (1.41)
9	RNA polymerase II cis-regulatory region sequence-specific DNA binding	53 (1.36)
10	Enzyme binding	50 (1.29)

Table-10: Top 10 most abundant biological process terms

Rank	Gene Ontology biological process terms	Count (Proportion %)
1	Positive regulation of transcription by RNA polymerase II	98 (1.09)
2	Negative regulation of transcription by RNA polymerase II	80 (0.89)
3	Regulation of transcription by RNA polymerase II5	69 (0.77)

4	Positive regulation of transcription, DNA-templated	65 (0.72)
5	Signal transduction	57 (0.63)
6	Negative regulation of transcription, DNA-templated	53 (0.59)
7	Negative regulation of apoptotic process	47 (0.52)
8	DNA repair	46 (0.51)
9	Cell division	41 (0.46)
10	Cellular response to DNA damage stimulus	40 (0.45)

Information regarding subcellular localization, cellular functions and biological activity of the proteins used in this study was obtained with the help of Gene Ontology terms. There are 3 kinds of Gene Ontology terms – cellular component, molecular function and biological process. All the proteins from the dataset were mapped to their respective Gene Ontology terms and these terms were sorted in a descending order of their abundance (Tables-8 to 10).

SUMOylated proteins mostly localize to nucleus. However, there are a few proteins that localize to cytoplasmic organelles or plasma membranes too (Table-8). Majority of the SUMOylated proteins bind nucleic acids such as DNA / RNA / ATP (Table-9). SUMOylated proteins such as zinc finger proteins are also known to bind metal ions. Biological processes such as transcription regulation, cell division, signal transduction and DNA repair have been linked to SUMOylation (Table-10).

5.3.2 Predictions made using residue contacts

The sampling method was applied to every lysine in all the target proteins. Clash-free poses were obtained for around half of all the SUMOylated and non-SUMOylated lysines (Table-11). For the remaining lysines, clash-free poses could not be obtained because either the main chain atoms of the target protein and ubc9 had collisions. This could be either due to unfavorable phi and psi angles of the lysine residues or the target protein conformation may not be optimal for binding ubc9.

Table-11: Overview of clash-free poses generated by the sampling method

	SUMOylated lysines (7432)	Non-SUMOylated lysines (27874)
Lysines with clash-free poses	4006	13022
Lysines without clash-free poses	3426	14852

Table-12: Proportion of lysines with clash-free poses that conform to consensus motif

	K-x-(E/D)	(E/D)-x-K	Non-consensus (E/D)
Clash-free SUMOylated lysines (4006)	42	29	613
Clash-free non-SUMOylated lysines (13022)	51	36	1949

A vast majority of lysines with clash-free poses do not conform to consensus motif (Table-12). There are 613 SUMOylated and 1949 non-SUMOylated lysines that have poses wherein an E/D binds positively charged patch on ubc9. But the E/D residue is not at +2 / -2 position with respect to the sequential position of lysine of interest. These E/D residues are referred to as non-consensus.

The Apriori algorithm generated res-pair sets varying in size from 1 to 18. Predictions were made independently for every res-pair set according to their size. Thus, predictions were made for all size-1 sets, size-2 sets and so on. The best predictions in terms of MCC were obtained for size-3 res-pair sets (Table-12). Predictions for size-3 sets were made by varying the cutoff from 1 to 185. Here cutoff refers to the number of size-3 res-pair set present in a given conformational pose. All poses having more res-pair sets than the cutoff were chosen as positive predictions or else they were marked as negative prediction.

The prediction method described here achieved a sensitivity = 27%, specificity = 98%, accuracy = 81% and MCC = 0.4 (Table-13). Our method has higher specificity than sensitivity. This can be attributed to the higher number of non-

SUMOylated poses (13022) than SUMOylated poses (4006). Future prediction tools can overcome this issue by under-sampling non-SUMOylated poses or over-sampling SUMOylated poses.

Table-13: Overview of predictions made for size-3 res-pair sets

Statistical parameter	Value
True positives	1091
True negatives	12763
False positives	259
False negatives	2915
Sensitivity	0.272
Specificity	0.98
Accuracy	0.814
F1 score	0.407
MCC	0.396

Table-14: Proportion of predicted lysines that conform to consensus motif

	K-x-(E/D)	(E/D)-x-K	Non-consensus (E/D)
True positives	13	7	301
False positives	1	0	90

Majority of lysines predicted using size-3 res-pair sets do not conform to consensus motif (Table-14). This trend is similar to the trend observed for lysines with clash-free poses (Table-12). An interesting observation is that true positives have more consensus lysines than false positives (20 versus 1).

Table-15: Top 10 size-3 res-pair sets that show maximum enrichments

Res-pair set	Occurrence in true positives	Occurrence in false positives
74;LEU, 88;LEU, 91;LEU	8	2
131;TYR, 129;TYR, 135;TYR	7	5
139;GLU, 136;ARG, 133;ARG	7	6

98;ALA, 99;GLN, 98;GLN	6	2
72;ARG, 99;ARG, 65;ASN	6	1
99;ARG, 71;ARG, 76;GLU	6	3
87;MET, 91;ASP, 74;ASP	6	4
76;ASP, 87;TRP, 89;ASP	5	1
131;SER, 129;TYR, 88;TYR	5	1
88;LEU, 129;LEU, 91;LEU	5	4

There were 7826 size-3 res-pair sets that showed positive enrichment in SUMOylated poses than non-SUMOylated poses. Out of these, top 10 res-pair sets in terms of their enrichment values are given above (Table-15). There are 2 sets of particular interest. These are 87;MET, 91;ASP, 74;ASP and 76:ASP, 87;TRP, 89;ASP. Both these sets represent contacts between an aspartate residue in target protein and positively charged patch on ubc9.

Secondary structure environment of SUMOylated lysines was determined using write_data() function in MODELLER. SUMOylation targets lysine residues in all secondary structures such as alpha helices, beta sheets or coils (Table-16). The sampling method and res-pair based predictions were able to detect SUMOylated lysines irrespective of their secondary structures (Table-16).

Table-16: Overview of secondary structures of SUMOylated lysines

Secondary structure	All SUMOylated lysines in the dataset (7432)	SUMOylated lysines with clash-free poses (4006)	True positives (1091)
Beta strand	1300	481	155
Loop	3095	1845	422
Alpha helix	2988	1658	504
Kink	49	22	10

Table-17: Secondary structures of SUMOylated lysines in consensus motif

Secondary structure	All SUMOylated consensus motif lysines (2556)	SUMOylated consensus motif lysines with clash-free poses (71)	Consensus motif lysines in True positives (20)

Beta strand	443	15	4
Loop	1087	50	16
Alpha helix	1011	6	0
Kink	15	0	0

SUMOylated lysines that follow a consensus motif occur in all kinds of secondary structures (Table-17). Sampling and prediction methods preferred lysines in a coil / loop. This is in agreement with lysine 524 in RanGAP1, which also happens to be in a located in a loop.

5.4 Discussion

All the existing computational methods suffer from the major drawback that they are biased in the favor of the ψ -K-x-(E/D) consensus motif. However, the consensus motif accounts for half of all the known SUMOylation sites. Thus, all the existing prediction tools are inefficient at predicting other half of the SUMOylation sites that do not follow the consensus motif. In order to make more robust predictions than existing tools, the present study proposed and demonstrated a method that uses protein 3-D structures rather than protein sequences to predict SUMOylation sites. The method described here achieved an accuracy of 81% and Matthews' correlation coefficient of 0.4.

SUMOylation is a dynamic post translational modification. SUMOylation can happen to proteins varying in size from 100 amino acids to more than 10000 amino acids (Figure-5B). The modification could target either one lysine in a given protein or more than 20 lysines (Figure-5A). Proteins belonging to different folds (CATH superfamilies) are targeted by the modification (Table-7). SUMOylation targets co-localize to either nucleus of a cell or to different cytoplasmic organelles and even the plasma membrane (Table-8). The target proteins could possibly bind DNA / RNA / ATP (Table-9) and might be involved in regulation of transcription activity, cell division, DNA repair or signal transduction (Table-10). SUMOylated lysines can occur in any of the 3 secondary structure environments such as alpha helices, beta sheets or loops / coils (Tables-16 and 17).

At present, the Protein Data Bank has structural information for only one ubc9-target complex. Almost all of the proteins used in this study were not bound to ubc9. Hence, the conformation used for sampling method may not necessarily be optimal for binding ubc9. Apart from this, factors such as crystal contacts could also influence protein conformations [55]. Hence, the poses generated by the sampling method have to be analyzed with caution. As more structural information becomes available for these interactions, more robust structure based prediction tools can be developed. Sampling different conformational poses and scoring those poses are two important aspects of any general protein-protein docking tool. In the present study, those two concepts were used for studying ubc9-target protein interactions. In cases of SUMOylated proteins with unknown 3-D structures, information from Alphafold models could be used [25]. In addition, future prediction tools can achieve improved accuracy by taking into account information about SUMO E3 ligases.

Chapter 6: Conclusions and Future prospects

Research conducted around the globe over the past two and a half decades has revealed biological implications of SUMOylation. Mutation of SUMOylated lysines from target proteins often results in diseases. The present thesis has proposed and demonstrated the utility of different computational approaches to analyze and predict SUMOylated lysines. These methods can be applied to study any post translational modification. However the present thesis focuses on SUMOylation. Information obtained from recent mass-spectrometry based proteomics experiments was used as input for the computational methods discussed here.

The sequence based approach begins with identification of homologous proteins with conserved SUMOylated lysines across different organisms. Specifically in this case, we started with human and mouse proteins having known SUMOylated lysines. We used the protein sequence alignment tool PSIBLAST to identify homologs from fruit fly *Drosophila melanogaster* that contain conserved lysines. In order to gain confidence in our predictions, we compared our list of fly homologs with other fly SUMO proteomics experiments that were able to identify fly SUMOylated proteins but not the modified lysines therein. Apart from the list of homologous proteins, we also obtained three kinds of analysis. First kind of analysis was done to find out amino acid patterns involving these SUMOylated lysines in their local sequence environment. This helped us detect sequence motifs involving these lysines. Second kind of analysis was done to find out which protein families tend to get more SUMOylated as compared to others. Third kind of analysis was done to find out preferred biological functions of the newly identified homologs. This analysis was done using Gene ontology annotations of these proteins.

The results obtained from the sequence based approach described above will be made available to the scientific community in the form of a database called

SUMO-ON-THE-FLY. Geneticists will be able to experimentally validate the predictions made therein, thus enriching our understanding of fly proteins.

To the best of our knowledge, all the available SUMOylation site prediction tools make use protein sequence information. However, candidate lysines are selected as a result of protein-protein interactions between enzyme (ubc9) and its target (substrate) proteins. Hence, we proposed a novel structure based prediction tool. The major hurdle for our efforts here was that the Protein Data Bank contains only one structure of the enzyme-target complex as of date. We circumvented this drawback by designing a new special docking technique which we refer to as sampling method. In addition, we designed a scoring method that can discriminate between SUMOylatable and non-SUMOylatable lysines on the basis of residue contacts between ubc9 and the target proteins. Our method achieved an accuracy of 81% and a Matthews' Correlation Coefficient of 0.4. Thus, we have set up the stage for the development of future structure based tools to predict post translational modification sites.

Chapter 7: Publications

1. Deshmukh P, Markande S, Fandade V, **Ramtirtha Y**, Madhusudhan MS, Joseph J. The miRISC component AGO2 has multiple binding sites for Nup358 SUMO-interacting motif. *Biochem Biophys Res Commun*. 2021 Jun 4;556:45-52. doi: 10.1016/j.bbrc.2021.03.140. Epub 2021 Apr 7. PMID: 33838501
2. Sahoo M, Gaikwad S, Khupekar D, Ashok M, Helen M, Yadav S, Singh A, Magre I, Deshmukh P, Dhanvijay S, Sahoo P, **Ramtirtha Y**, Madhusudhan M, Gayathri P, Seshadri V, Joseph J. “Nup358 binds to AGO proteins through its SUMO-interacting motifs and promotes the association of target mRNA with miRISC.” *EMBO reports* vol. 18,2 (2017): 241-263. doi:10.15252/embr.201642386
3. T. R. Kanitkar, N. Sen, S. Nair , N. Soni, K. Amritkar, **Y. Ramtirtha**, and M. Madhusudhan, “Methods for Molecular Modelling of Protein Complexes,” *Methods in Molecular Biology* (Clifton, NJ), vol. 2305, pp. 53–80, 2021.
4. **Y. Ramtirtha**, M. S. Madhusudhan, Prediction of SUMOylation targets in *Drosophila melanogaster*, *BioRxiv*, <https://www.biorxiv.org/content/10.1101/2022.08.19.504577v1>
5. **Y. Ramtirtha**, M. S. Madhusudhan, 3-D structure based prediction of SUMOylation sites, *BioRxiv*, <https://www.biorxiv.org/content/10.1101/2022.08.19.504594v1>

Chapter 8: Appendix

1. Study of interaction between human Argonaute-2 and Nup358

This project was done in collaboration with the research group of Dr. Jomon Joseph, NCCS Pune. Their experiments showed that Argonaute-2 bound Nup358, a protein of nuclear pore complex. Argonaute has many domains that look structurally similar to SUMO. Nup358 is known to bind SUMO too. Hence, we used a structure superimposition tool called CLICK to study structural similarity between SUMO and Argonaute. The results from this work are a part of publications-1 and 2 given above.

2. Molecular dynamics simulations of 14-3-3 proteins

This project was done in collaboration with Dr. Prasanna Venkatraman, ACTREC, Navi Mumbai. 14-3-3 proteins form homodimers that bind phosphopeptides and many small ligand molecules. Our collaborators have data to show that 14-3-3 proteins also bind ATP. So, Neelesh Soni (a senior from the lab) modeled the complex of 14-3-3 with ATP. We carried out molecular dynamics simulations of zeta and gamma isoforms of 14-3-3 homodimers in apo and holo forms using the GROMACS software. The simulations were carried out for 150 ns and the GROMOS force field was used.

Chapter 9: Supporting Information

All the supporting information related to the present thesis has been uploaded to Github. Here is the link to the zip file containing all the raw data relevant to Chapter 3- <https://github.com/yogendra-bioinfo/homology-based-SUMOylation-prediction.git>. Here is the link to the zip file containing all the raw data relevant to Chapter 5 - <https://github.com/yogendra-bioinfo/structure-based-SUMOylation-prediction.git>. Each folder contains a README file that explains the results present in all the files given therein.

Chapter 10: Bibliography

- [1] R. Mahajan, C. Delphin, T. Guan, L. Gerace, F. Melchior, A Small Ubiquitin-Related Polypeptide Involved in Targeting RanGAP1 to Nuclear Pore Complex Protein RanBP2, *Cell*. 88 (1997) 97–107. [https://doi.org/10.1016/S0092-8674\(00\)81862-0](https://doi.org/10.1016/S0092-8674(00)81862-0).
- [2] A. Flotho, F. Melchior, Sumoylation : A Regulatory Protein Modification in Health and Disease, (n.d.). <https://doi.org/10.1146/annurev-biochem-061909-093311>.
- [3] S. Ramazi, J. Zahiri, Post-translational modifications in proteins: Resources, tools and prediction methods, *Database*. 2021 (2021) 1–20. <https://doi.org/10.1093/database/baab012>.
- [4] U. Sahin, H. de Thé, V. Lallemand-Breitenbach, Sumoylation in Physiology, Pathology and Therapy, *Cells*. 11 (2022) 1–24. <https://doi.org/10.3390/cells11050814>.
- [5] N.E. Pellegrino, A. Guven, K. Gray, P. Shah, G. Kasture, M.D. Nastke, A. Thakurta, S. Gesta, V.K. Vishnudas, N.R. Narain, M.A. Kiebish, The Next Frontier: Translational Development of Ubiquitination, SUMOylation, and NEDDylation in Cancer, *Int. J. Mol. Sci*. 23 (2022). <https://doi.org/10.3390/ijms23073480>.
- [6] Y. Fan, X. Li, L. Zhang, Z. Zong, F. Wang, J. Huang, L. Zeng, C. Zhang, H. Yan, L. Zhang, F. Zhou, SUMOylation in Viral Replication and Antiviral Defense, *Adv. Sci*. 9 (2022) 1–14. <https://doi.org/10.1002/advs.202104126>.
- [7] A. Bateman, M.J. Martin, S. Orchard, M. Magrane, R. Agivetova, S. Ahmad, E. Alpi, E.H. Bowler-Barnett, R. Britto, B. Bursteinas, H. Bye-A-Jee, R. Coetzee, A. Cukura, A. Da Silva, P. Denny, T. Dogan, T.G. Ebenezer, J. Fan, L.G. Castro, P. Garmiri, G. Georghiou, L. Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, P. Jokinen, V. Joshi, D. Jyothi, A. Lock, R. Lopez, A. Luciani, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, M. Menchi, A. Mishra, K. Moulang, A. Nightingale, C.S. Oliveira, S. Pundir, G. Qi, S. Raj, D. Rice, M.R. Lopez, R. Saidi, J. Sampson, T. Sawford, E. Speretta, E. Turner, N. Tyagi, P. Vasudev, V. Volynkin, K. Warner, X. Watkins, R. Zaru, H. Zellner, A. Bridge, S.

- Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.C. Blatter, J. Bolleman, E. Boutet, L. Breuza, C. Casals-Casas, E. de Castro, K.C. Echioukh, E. Coudert, B. Cuche, M. Doche, D. Dornevil, A. Estreicher, M.L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, G. Keller, A. Kerhornou, V. Lara, P. Le Mercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T.B. Neto, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, M. Pozzato, M. Pruess, C. Rivoire, C. Sigrist, K. Sonesson, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, C.H. Wu, C.N. Arighi, L. Arminski, C. Chen, Y. Chen, J.S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D.A. Natale, K. Ross, C.R. Vinayaka, Q. Wang, Y. Wang, L.S. Yeh, J. Zhang, UniProt: The universal protein knowledgebase in 2021, *Nucleic Acids Res.* 49 (2021) D480–D489. <https://doi.org/10.1093/nar/gkaa1100>.
- [8] Z. Li, S. Li, M. Luo, J.-H. Jhong, W. Li, L. Yao, Y. Pang, Z. Wang, R. Wang, R. Ma, J. Yu, Y. Huang, X. Zhu, Q. Cheng, H. Feng, J. Zhang, C. Wang, J.B.-K. Hsu, W.-C. Chang, F.-X. Wei, H.-D. Huang, T.-Y. Lee, dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications, *Nucleic Acids Res.* 50 (2021) D471–D479. <https://doi.org/10.1093/nar/gkab1017>.
- [9] H. Xu, J. Zhou, S. Lin, W. Deng, Y. Zhang, Y. Xue, PLMD: An updated data resource of protein lysine modifications., *J. Genet. Genomics.* 44 (2017) 243–250. <https://doi.org/10.1016/j.jgg.2017.03.007>.
- [10] P. V Hornbeck, B. Zhang, B. Murray, J.M. Kornhauser, V. Latham, E. Skrzypek, PhosphoSitePlus, 2014: mutations, PTMs and recalibrations, *Nucleic Acids Res.* 43 (2014) D512–D520. <https://doi.org/10.1093/nar/gku1267>.
- [11] I.A. Hendriks, A.C.O. Vertegaal, A comprehensive compilation of SUMO proteomics, *Nat. Rev. Mol. Cell Biol.* 17 (2016) 581–595. <https://doi.org/10.1038/nrm.2016.81>.
- [12] I. a Hendriks, R.C.J. D'Souza, B. Yang, M. Verlaan-de Vries, M. Mann, A.C.O. Vertegaal, Uncovering global SUMOylation signaling networks in a site-specific manner, *Nat. Struct. Mol. Biol.* 21 (2014) 927–936. <https://doi.org/10.1038/nsmb.2890>.
- [13] I.A. Hendriks, D. Lyon, D. Su, N.H. Skotte, J.A. Daniel, L.J. Jensen, M.L. Nielsen, Site-specific characterization of endogenous SUMOylation across

- species and organs, *Nat. Commun.* 9 (2018). <https://doi.org/10.1038/s41467-018-04957-4>.
- [14] Q. Zhao, Y. Xie, Y. Zheng, S. Jiang, W. Liu, W. Mu, Z. Liu, Y. Zhao, Y. Xue, J. Ren, GPS-SUMO: A tool for the prediction of sumoylation sites and SUMO-interaction motifs, *Nucleic Acids Res.* 42 (2014) 325–330. <https://doi.org/10.1093/nar/gku383>.
- [15] J. Ren, X. Gao, C. Jin, M. Zhu, X. Wang, A. Shaw, L. Wen, X. Yao, Y. Xue, Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0, *Proteomics.* 9 (2009) 3409–3412. <https://doi.org/10.1002/pmic.200800646>.
- [16] G. Beauclair, A. Bridier-Nahmias, J.-F. Zagury, A. Saïb, A. Zamborlini, JASSA: a comprehensive tool for prediction of SUMOylation sites and SIMs., *Bioinformatics.* 31 (2015) 3483–3491. <https://doi.org/10.1093/bioinformatics/btv403>.
- [17] J. Xu, Y. He, B. Qiang, J. Yuan, X. Peng, X.-M. Pan, A novel method for high accuracy sumoylation site prediction from protein sequences., *BMC Bioinformatics.* 9 (2008) 8. <https://doi.org/10.1186/1471-2105-9-8>.
- [18] Y.Z. Chen, Z. Chen, Y.A. Gong, G. Ying, SUMOhydro: A novel method for the prediction of sumoylation sites based on hydrophobic properties, *PLoS One.* 7 (2012) 1–8. <https://doi.org/10.1371/journal.pone.0039195>.
- [19] A. Ijaz, SUMOhunt: Combining Spatial Staging between Lysine and SUMO with Random Forests to Predict SUMOylation, *ISRN Bioinforma.* 2013 (2013) 1–11. <https://doi.org/10.1155/2013/671269>.
- [20] A. Yavuz, O. Sezerman, Predicting sumoylation sites using support vector machines based on various sequence features, conformational flexibility and disorder, *BMC Genomics.* 15 (2014) S18. <https://doi.org/10.1186/1471-2164-15-S9-S18>.
- [21] C.-C. Chang, C.-H. Tung, C.-W. Chen, C.-H. Tu, Y.-W. Chu, SUMOgo: Prediction of sumoylation sites on lysines by motif screening models and the effects of various post-translational modifications, *Sci. Rep.* 8 (2018) 15512. <https://doi.org/10.1038/s41598-018-33951-5>.
- [22] A. Dehzangi, Y. López, G. Taherzadeh, A. Sharma, T. Tsunoda, SumSec: Accurate Prediction of Sumoylation Sites Using Predicted Secondary Structure., *Molecules.* 23 (2018).

- <https://doi.org/10.3390/molecules23123260>.
- [23] A. Sharma, A. Lysenko, Y. López, A. Dehzangi, R. Sharma, H. Reddy, A. Sattar, T. Tsunoda, HseSUMO: Sumoylation site prediction using half-sphere exposures of amino acids residues, *BMC Genomics*. 19 (2019) 982. <https://doi.org/10.1186/s12864-018-5206-8>.
- [24] J. Jia, L. Zhang, Z. Liu, X. Xiao, K.-C. Chou, pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC, *Bioinformatics*. 32 (2016) 3133–3141. <https://doi.org/10.1093/bioinformatics/btw387>.
- [25] J. Zahiri, S. Ramazi, Computational Prediction of Proteins Sumoylation: A Review on the Methods and Databases, *J. Nanomedicine Res.* 3 (2016). <https://doi.org/10.15406/jnmr.2016.03.00068>.
- [26] Y.-W. Zhao, S. Zhang, H. Ding, Recent Development of Machine Learning Methods in Sumoylation Sites Prediction, *Curr. Med. Chem.* 29 (2021) 894–907. <https://doi.org/10.2174/0929867328666210915112030>.
- [27] M. Nie, Y. Xie, J.A. Loo, A.J. Courey, Genetic and proteomic evidence for roles of Drosophila SUMO in cell cycle control, Ras signaling, and early pattern formation, *PLoS One*. 4 (2009). <https://doi.org/10.1371/journal.pone.0005905>.
- [28] M. Handu, B. Kaduskar, R. Ravindranathan, A. Soory, R. Giri, V.B. Elango, H. Gowda, G.S. Ratnaparkhi, SUMO-enriched proteome for drosophila innate immune response, *G3 Genes, Genomes, Genet.* 5 (2015) 2137–2154. <https://doi.org/10.1534/g3.115.020958>.
- [29] L. Pirone, W. Xolalpa, J.O. Sigursson, J. Ramirez, C. Pérez, M. González, A.R. De Sabando, F. Elortza, M.S. Rodriguez, U. Mayor, J. V. Olsen, R. Barrio, J.D. Sutherland, A comprehensive platform for the analysis of ubiquitin-like protein modifications using in vivo biotinylation, *Sci. Rep.* 7 (2017) 1–17. <https://doi.org/10.1038/srep40756>.
- [30] Q. Zhao, Y. Xie, Y. Zheng, S. Jiang, W. Liu, W. Mu, Z. Liu, Y. Zhao, Y. Xue, J. Ren, GPS-SUMO: A tool for the prediction of sumoylation sites and SUMO-interaction motifs, *Nucleic Acids Res.* 42 (2014) 325–330. <https://doi.org/10.1093/nar/gku383>.
- [31] J. Ren, X. Gao, C. Jin, M. Zhu, X. Wang, A. Shaw, L. Wen, X. Yao, Y. Xue,

- Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0, *Proteomics*. 9 (2009) 3409–3412.
<https://doi.org/https://doi.org/10.1002/pmic.200800646>.
- [32] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs., *Nucleic Acids Res.* 25 (1997) 3389–3402.
<https://doi.org/10.1093/nar/25.17.3389>.
- [33] A.A. Schäffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E. V. Koonin, S.F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res.* 29 (2001) 2994–3005.
<https://doi.org/10.1093/nar/29.14.2994>.
- [34] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, in: *Proc. 20th Int. Conf. Very Large Data Bases*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994: pp. 487–499.
<http://dl.acm.org/citation.cfm?id=645920.672836>.
- [35] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics*. 26 (2010) 680–682. <https://doi.org/10.1093/bioinformatics/btq003>.
- [36] W. Li, L. Jaroszewski, A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases , *Bioinformatics*. 18 (2002) 77–82. <https://doi.org/10.1093/bioinformatics/18.1.77>.
- [37] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*. 22 (2006) 1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>.
- [38] W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases , *Bioinformatics*. 17 (2001) 282–283. <https://doi.org/10.1093/bioinformatics/17.3.282>.
- [39] A. Sali, T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.* 234 (1993) 779–815.
<https://doi.org/10.1006/jmbi.1993.1626>.
- [40] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P.

- Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with {NumPy}, *Nature*. 585 (2020) 357–362. <https://doi.org/10.1038/s41586-020-2649-2>.
- [41] H. Chen, P.C. Boutros, VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R, *BMC Bioinformatics*. 12 (2011) 35. <https://doi.org/10.1186/1471-2105-12-35>.
- [42] R Core Team, R: A Language and Environment for Statistical Computing, (2018). <https://www.r-project.org/>.
- [43] A. Larkin, S.J. Marygold, G. Antonazzo, H. Attrill, G. dos Santos, P. V Garapati, J.L. Goodman, L.S. Gramates, G. Millburn, V.B. Strelets, C.J. Tabone, J. Thurmond, F. Consortium, FlyBase: updates to the *Drosophila melanogaster* knowledge base, *Nucleic Acids Res.* 49 (2020) D899–D907. <https://doi.org/10.1093/nar/gkaa1026>.
- [44] Q. Zhao, Y. Xie, Y. Zheng, S. Jiang, W. Liu, W. Mu, Z. Liu, Y. Zhao, Y. Xue, J. Ren, GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs, *Nucleic Acids Res.* 42 (2014) W325–W330. <https://doi.org/10.1093/nar/gku383>.
- [45] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, *Nucleic Acids Res.* 28 (2000) 235–242. <https://doi.org/10.1093/nar/28.1.235>.
- [46] D. Reverter, C.D. Lima, Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex., *Nature*. 435 (2005) 687–692. <https://doi.org/10.1038/nature03588>.
- [47] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera - A visualization system for exploratory research and analysis, *J. Comput. Chem.* 25 (2004) 1605–1612. <https://doi.org/10.1002/jcc.20084>.
- [48] J.M. Dana, A. Gutmanas, N. Tyagi, G. Qi, C. O’Donovan, M. Martin, S. Velankar, SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins, *Nucleic Acids Res.* 47 (2018) D482–D489. <https://doi.org/10.1093/nar/gky1114>.
- [49] S. Velankar, J.M. Dana, J. Jacobsen, G. van Ginkel, P.J. Gane, J. Luo, T.J. Oldfield, C. O’Donovan, M.-J. Martin, G.J. Kleywegt, SIFTS: Structure

- Integration with Function, Taxonomy and Sequences resource, *Nucleic Acids Res.* 41 (2012) D483–D489. <https://doi.org/10.1093/nar/gks1258>.
- [50] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, 2016. <https://ggplot2.tidyverse.org>.
- [51] K.S. Arun, T.S. Huang, S.D. Blostein, Least-Squares Fitting of Two 3-D Point Sets, *IEEE Trans. Pattern Anal. Mach. Intell.* 9 (1987) 698–700. <https://doi.org/10.1109/TPAMI.1987.4767965>.
- [52] M. V Shapovalov, R.L.J. Dunbrack, A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions., *Structure.* 19 (2011) 844–858. <https://doi.org/10.1016/j.str.2011.03.019>.
- [53] P.A. Wang, Junmei and Cieplak, Piotr and Kollman, How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?, *J. Comput. Chem.* 21 (2000) 1049–1074. [https://doi.org/https://doi.org/10.1002/1096-987X\(200009\)21:12<1049::AID-JCC3>3.0.CO;2-F](https://doi.org/https://doi.org/10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F).
- [54] I. Sillitoe, N. Bordin, N. Dawson, V.P. Waman, P. Ashford, H.M. Scholes, C.S.M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S.D. Lam, K. Berka, I.H. Varekova, R. Svobodova, J. Lees, C.A. Orengo, CATH: increased structural coverage of functional space, *Nucleic Acids Res.* 49 (2020) D266–D273. <https://doi.org/10.1093/nar/gkaa1079>.
- [55] P. Radivojac, V. Vacic, C. Haynes, R.R. Cocklin, A. Mohan, J.W. Heyen, M.G. Goebel, L.M. Iakoucheva, Identification, analysis, and prediction of protein ubiquitination sites., *Proteins.* 78 (2010) 365–380. <https://doi.org/10.1002/prot.22555>.
- [56] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, *Nature.* 596 (2021) 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.

