GPS-SUMO Manual

Prediction of SUMOylation Sites & SUMO-interaction Motifs

Version 1.0.1
05/10/2014

Author: Jian Ren & Yu Xue

Contact: Dr. Jian Ren, renjian.sysu@gmail.com; Dr. Yu Xue, xueyu@hust.edu.cn
The software is only free for academic research.
The latest version of GPS-SUMO software is available from http://sumosp.biocuckoo.org
Copyright (c) 2006-2014. The CUCKOO Workgroup. All Rights Reserved.
Index

STATEMENT ........................................................................................................................................ 2
INTRODUCTION ................................................................................................................................. 3
DOWNLOAD & INSTALLATION ........................................................................................................... 5
PREDICTION OF SUMO MODIFICATION ......................................................................................... 8
  A SINGLE PROTEIN SEQUENCE IN FASTA FORMAT ................................................................. 8
  MULTIPLE PROTEIN SEQUENCES IN FASTA FORMAT ............................................................. 11
THE USAGE OF SUMO DATABASE ..................................................................................................... 15
  SEARCH .......................................................................................................................................... 15
  ADVANCE SEARCH ......................................................................................................................... 18
  BROWSE ......................................................................................................................................... 20
  BLAST SEARCH .............................................................................................................................. 21
REFERENCES ..................................................................................................................................... 24
RELEASE NOTE ................................................................................................................................. 25
Statement

1. **Implementation.** The softwares of the CUCKOO Workgroup are implemented in JAVA (J2SE). Usually, both of online service and local stand-alone packages will be provided.

2. **Availability.** Our softwares are freely available for academic researches. For non-profit users, you can copy, distribute and use the softwares for your scientific studies. Our softwares are not free for commercial usage.

3. **GPS.** Previously, we used the GPS to denote our Group-based Phosphorylation Scoring algorithm. Currently, we are developing an integrated computational platform for post-translational modifications (PTMs) of proteins. We re-denote the GPS as Group-based Prediction Systems. This software is an indispensable part of GPS.

4. **Usage.** Our softwares are designed in an easy-to-use manner. Also, we invite you to read the manual before using the softwares.

5. **Updation.** Our softwares will be updated routinely based on users’ suggestions and advices. Thus, your feedback is greatly important for our future updation. Please do not hesitate to contact with us if you have any concerns.

6. **Citation.** Usually, the latest published articles will be shown on the software websites. We wish you could cite the article if the software has been helpful for your work.

7. **Acknowledgements.** The work of CUCKOO Workgroup is supported by grants from the National Basic Research Program (973 project) [2013CB933902, 2012CB911201, 2012CB910101]; National Natural Science Foundation of China [31171263, 81272578]; the Guangdong Natural Science Funds for Distinguished Young Scholar [S20120011335]; Zhujiang Nova Program of Guangzhou [2011J2200042]; Program of International S&T Cooperation [2014DFB30020].
Introduction

Among the many protein post-translational modifications, sumoylation acts as a crucial biochemical process in the regulation of a variety of important biological functions. By specifically attaching a SUMO protein to a substrate, protein sumoylation could regulate multiple biochemical properties of protein target like the stability, activity, intracellular localization and protein interactions (1-3). Thousands of studies uncovered that sumoylation is essential for a series of cellular processes, including DNA damage recovery, gene expression, chromosomal integrity as well as nuclear protein assemblies (4,5). In addition, protein sumoylation has shown to be intimately correlated with human diseases such as Alzheimer’s disease (AD) (6), Parkinson’s disease (PD) (7), viral infections (8), cardiac disease (9,10) and cancers (11).

Recently progresses revealed a class of SUMO-interaction motifs (SIMs) or SUMO-interacting motifs (SIMs), which mediate non-covalent interaction between SUMO and other proteins. The hydrophobic core of V/I-X-V/I-V/I or similar motifs provide additional specificity for sumoylation, and generate an interface for protein-protein interaction. However, such a motif will generate a huge number of potential hits for proteomic survey, which might be difficult for ambiguously experimental verifications. To data, there were only dozens of SIMs experimentally identified in proteins. In this regard, an accurate and efficient predictor is in urgent need for further experimental manipulation.

In this work, we reported an update of SUMOsp2.0 and renamed it to GPS-SUMO by mainly adding a novel SIM prediction feature. We first manually collected 151 known SIMs in 80 proteins from scientific literature. A new generation GPS (Group-based Prediction System) algorithm integrated with PSO (Particle Swarm Optimization) method was employed for predictor training. Due to the data limitation, only the leave-one-out validation was carried out to evaluate the prediction performance, with the sensitivity (Sn) of 92.7% and the specificity (Sp) of 95.0%. Additionally, to cover classic sumoylation sites prediction, the improved GPS algorithm and latest dataset were applied (986 sumoylation sites in 545 proteins). As the first computational tool for prediction of both SIMs and sumoylation sites, the web service of GPS-SUMO were implemented in JAVA and PHP, which is freely available at http://sumosp.biocuckoo.org/.
### GPS-SUMO 1.0 User Interface

#### Predicted Sites

<table>
<thead>
<tr>
<th>ID</th>
<th>Position</th>
<th>Peptide</th>
<th>Score</th>
<th>Cutoff</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIN04</td>
<td>69</td>
<td><strong>SHVAAAPKLEIS</strong></td>
<td>18120</td>
<td>16</td>
<td>SUMOylation</td>
</tr>
<tr>
<td>GIN04</td>
<td>99</td>
<td><strong>YLKCGMAIEHFPS</strong></td>
<td>30573</td>
<td>16</td>
<td>SUMOylation</td>
</tr>
<tr>
<td>GIN04</td>
<td>192-198</td>
<td><strong>VERITQEN ILYER DTIEFSD</strong></td>
<td>63626</td>
<td>59.29</td>
<td>SUMOylation</td>
</tr>
<tr>
<td>RNF111</td>
<td>15</td>
<td><strong>TEKULTLEKVQERIK</strong></td>
<td>40499</td>
<td>38.825</td>
<td>SUMOylation</td>
</tr>
<tr>
<td>RNF111</td>
<td>19</td>
<td><strong>YLKLKSMVIGIEDPA</strong></td>
<td>31495</td>
<td>16</td>
<td>SUMOylation</td>
</tr>
<tr>
<td>RNF111</td>
<td>110</td>
<td><strong>STYKQCYKERKILD</strong></td>
<td>37643</td>
<td>35.625</td>
<td>SUMOylation</td>
</tr>
<tr>
<td>RNF111</td>
<td>300-304</td>
<td><strong>CQHSDK TVDSY ASSTQY</strong></td>
<td>68235</td>
<td>59.29</td>
<td>SUMOylation</td>
</tr>
<tr>
<td>RNF111</td>
<td>220-220</td>
<td><strong>VSHS RVEIT VUESR</strong></td>
<td>59692</td>
<td>59.29</td>
<td>SUMOylation</td>
</tr>
<tr>
<td>RNF111</td>
<td>302-306</td>
<td><strong>LQNHAX VYKYE VUEHET</strong></td>
<td>7788</td>
<td>59.29</td>
<td>SUMOylation</td>
</tr>
</tbody>
</table>

#### Enter sequence(s) in FASTA format

```
MTKQNYLCLLNGGGLNDEGEEVLTALPHELRLGKWVHNENFHELLESITPEVGCYNHHELENLRRRGKASAGWGNNERRRLSDDLRCXLGMENKEFPMLSEKFRPGEFPL
RNF111
RGGTVTFLKFCTLQDNSEGKAPTHQGKLLHPFQAQPGWHSNOKGQEFHLCDDGQGESEMQMNQGEOQEGLQYYRQKSQKHPFYYKNGVHENGQLGF
```

**GPS-SUMO 1.0 User Interface**
Download & Installation

The GPS-SUMO was implemented in JAVA (J2SE), and could support three major Operating Systems (OS), including Windows, Linux/Unix or Mac OS X systems. Both of online web service and local stand-alone packages are available from: http://sumosp.biocuckoo.org/online.php. We recommend that users could download the latest release.

Please choose the proper package to download. After downloading, please double-click on the software package to begin installation, following the user prompts through the installation. And snapshots of the setup program for windows are shown below:
Select Destination Directory
Where should GPS-SUMO 1.0 be installed?

Select the folder where you would like GPS-SUMO 1.0 to be installed, then click Next.

Destination directory
Program Files\Group-based Prediction System\GPS-SUMO

Required disk space: 143.8 MB
Free disk space: 96,606 MB

Installing
Please wait while Setup installs GPS-SUMO 1.0 on your computer.
Finally, please click on the Finish button to complete the setup program.
Prediction of SUMO modification

A single protein sequence in FASTA format

The following steps show you how to use the GPS-SUMO 1.0 to predict SUMO modified sites for a single protein sequence in FASTA format.

1) Firstly, please use “Ctrl+C & Ctrl+V” (Windows & Linux/Unix) or “Command+C & Command+V” (Mac) to copy and paste your sequence into the text form of GPS-SUMO 1.0.

Note: for a single protein, the sequence without a name in raw format is also OK. However, for multiple sequences, the name of each protein should be presented.
(2) Choose a Thresholds of Sumoylation and SUMO-interaction that you need, the default cut-off is Medium. Note that you can choose “None” to exclude the prediction of a corresponding modification type.

(3) Click on the Submit button, then the predicted sumoylation sites/SUMO-interaction motif will be shown.

(4) Then please click on the RIGHT button in the prediction form. You can use the
“Select All” and “Copy Selected” to copy the selected results into Clipboard. Then please copy the results into a file, eg., an EXCEL file for further consideration. Also, you can choose “Export Prediction” to export the prediction results into a tab-delimited text file.

To visualize the predicted site please click on the “Visualize” menu.

Again, you can also click the “Export Prediction” in File menu to export the results.
Multiple protein sequences in FASTA format

For multiple protein sequences, there are two ways to use the GPS-SUMO 1.0.

A. Input the sequences into text form directly. (Num. of Seq \( \leq 2,000 \))
If the number of total protein sequences is not greater than 2,000, you can just use “Ctrl+C & Ctrl+V” (Windows & Linux/Unix) or “Command+C & Command+V” (Mac) to copy and paste your sequences into the text form of GPS-SUMO 1.0 for prediction.

B. Use Batch Predictor tool.
If the number of protein sequences is very large, eg., yeast or human proteome, please use the Batch Predictor. Please click on the “Batch Predictor” button in the Tools menu.
The following steps show you how to use it:

(1) Put protein sequences into one or several files (e.g., SC.fas, CE.fas, and etc) with FATSA format as below:

>protein1
XXXXXXXXXXXXXXX
XXXXXXXXX
>protein2
XXXXXXXXXXXXXXXXX...
>protein3
XXXXXXXXXXXXX
...
Most importantly, the name of each protein should be presented.

(2) Click on the Batch Predictor button and then click on the Add File button and add one or more protein sequence files in your hard disk.
Then the names of added files will be shown in the **Sequence File List**.

(3) The output directory of prediction results should also be defined. Please click on the >> button to specify the export fold.
(4) Please choose a proper threshold before prediction. Then please click on the Submit button, then the Batch Predictor begin to process all of the sequence files that have been added to the list. The result of prediction will be export to the Result Export Fold, and the name of result files will be shown in the Result File List.
The usage of SUMO database

In GPS-SUMO 1.0, a SUMO modification database was integrated. Please click on the “SUMO Database” button in Tool menu to launch the database.

Search

The SUMO database was designed in an easy-to-use manner. For simple search, users could input a SUMO ID with SUMOdb-XXXX-XXXXXX, a UniProt ID (P02768), protein/gene names/aliases (e.g., Serum albumin) or functions. Users could click the “Example” button one or several times to view the instances.

For example, users could input a protein/gene name/alias, e.g., G-protein, specify the “Any Field”, and then click on the “Submit” button to search the related information for this protein.
Then the information for G-protein will be shown in the “**Information**” form.

The FASTA sequence for this protein would also be shown in the “**Information**” form.
The searched results could be saved in HTML format.

Notably, you can click on the “Visualize” button to view the sumoylation site or SUMO-interaction motif (green in graph) in our DOG 2 package.
Advance search

The SUMO database supports three advance options, including advance search, browse, BLAST search. The Advance search option allows you to input up to three terms to find the information more specifically. The querying fields can be empty if fewer terms are needed.
First, users could click on the “Tools” button then click on the “Advance Search” button to open this option. By clicking the “Example” button, you can try an instance for usage. You can input DNA repair(Function), Histone (Protein Name), and 9606 (SUMOdb ID) for querying.

Then the result will be shown as follow:
Again, you can click on the “Visualize” button to open a schematic diagram for sumoylation site.

![Schematic diagram of protein SUMO modified sites](image)

**Browse**

The SUMO database supports the browse function. The Browse search allows users to view all entries in SUMO database.

First, users could click on the “Tools” button then click on the “Browse” button to visualize all SUMO proteins. Users could visualize any protein by click on the entries Listed in the “SUMO ID” form. Also, by clicking the “Visualize” button, a schematic diagram of protein SUMO modified sites will be shown.
Finally, all browsed results could be saved in HTML format.

**Blast search**

The SUMO database also supports the searching function by sequence alignment. The blastp program from NCBI BLAST+ packages was included in SUMO database. Users
could input one protein (not mRNA sequence) in FASTA or RAW format a time to search identical or homologous entries. First, users could click on the “Tools” button then click on the “Blast Search” button to open the Blast search window.

Then users could either click on the “Example” button in the Option form or directly input a protein sequence in FASTA or RAW format. Please note that only one protein is permitted a time. Then please click on the “Submit” button to search identical or
homologous entries. The E-value and species could be user-defined in the Option form.

Again, users could visualize any SUMO proteins by clicking on the entries listed in the “Matched list” form. And the results could be saved by clicking on the “Save” button in the Option form. Or you can click on the “Visualize” button to view the SUMO modified sites.
References

Release Note

1. Jan. 7th, 2008, the online service and the local stand-alone packages of SUMOsp 2.0 were released.
2. Jan. 29th, 2008, a bug was found that the version 2.0 couldn’t be properly used under non-English Operating Systems.
3. Feb. 16th, 2008, the version 2.0 manual was written and included in the packages.
4. Aug. 28th, 2008, **DOG (Domain Graph) 1.0** was integrated into SUMOsp 2.0.2 and a new function of visualizing the predicted sites was added. Feb. 2nd, 2009, SUMOsp version 2.0.3 was released. We moved the SUMOsp web server to a new website (http://sumosp.biocuckoo.org) and a new GPS logo was put into use.
5. Jul. 23rd, 2009, SUMOsp version 2.0.4 was released. Check for update function was added. DOG (Domain Graph) was updated to version 1.0.5.
6. Dec. 30th, 2013, by improving the prediction algorithm with the Particle Swarm Optimization (PSO) and adding the novel SUMO-interaction Motifs prediction feature, we developed an updated version of SUMOsp and renamed it as GPS-SUMO 1.0.