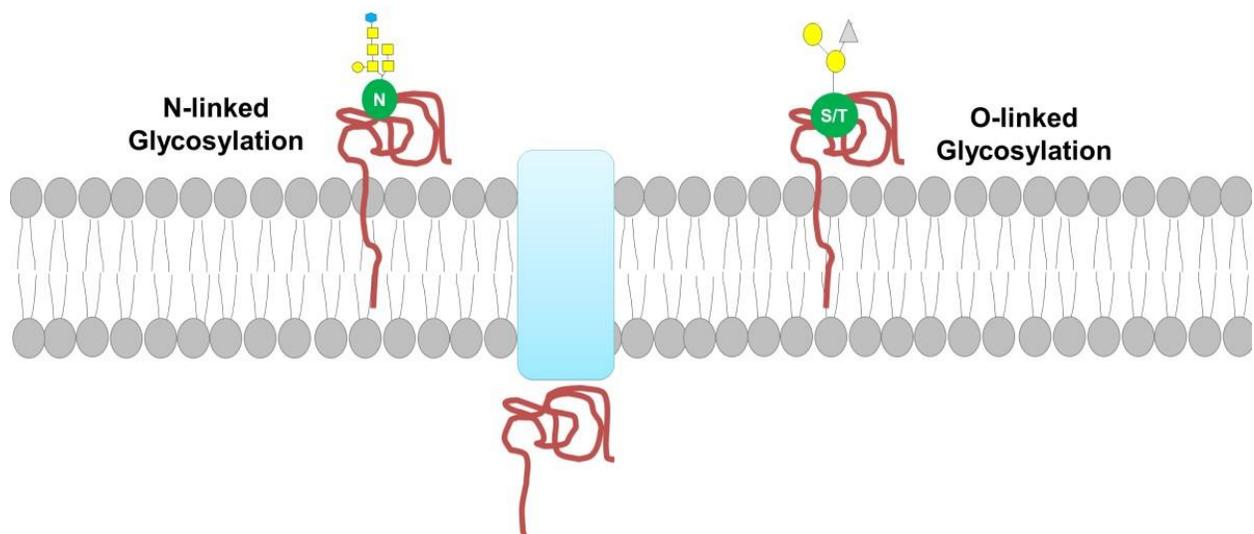


GlycoPP 2.0

User Manual Version 1.0

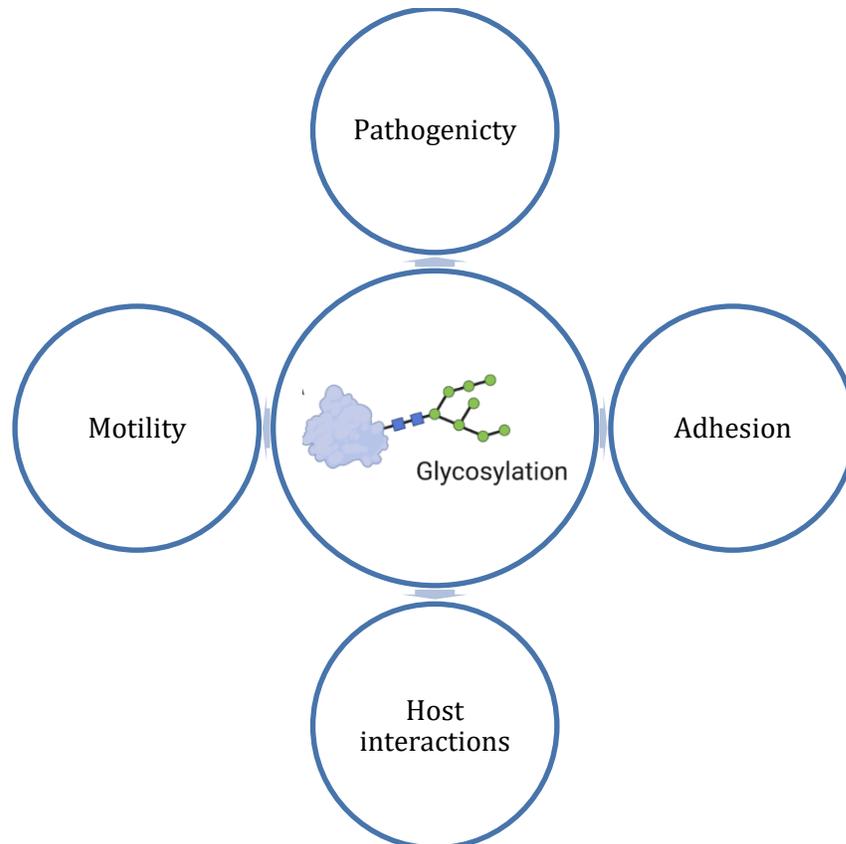


https://datascience.imtech.res.in/alkarao/glycopp_v2

The GlycoPP 2.0 web server is free to use without any registration/license conditions

1.0 Introduction

The glycosylated proteins play crucial roles in prokaryotes including motility, adhesion, host colonization, pathogenicity, immune regulation or elusion and thus identifying glycosylated residues in proteins will contribute significantly toward diagnostic and therapeutic development.



GlycoPP 2.0 is an upgraded version of the GlycoPP, which allows the users to predict N-linked and O-linked glycosylation sites in prokaryotic proteins without any registration/license conditions. The developed models available on this platform are trained on the experimentally identified N-linked and O-linked glycosylation sites in prokaryotes retrieved from ProGlycProt 2.0 database (<http://www.proglycprot.org/>).

The GlycoPP 2.0 is developed utilizing the open source, web-based platform Galaxy (<https://usegalaxy.org/>).

2.0 Accessing the GlycoPP 2.0 web server

The GlycoPP 2.0 can be accessed by the following ways-

2.1 GlycoPP 2.0 Home page- The user can access the homepage by following the link- <http://datascience.imtech.res.in/alkarao/glycopp2/index.html>. This page serves as a reference point to get an overview of the developed models and associated links including access to the Galaxy integration page. The home page is shown in Figure 1 below.

GlycoPP 2.0
A web-server for prokaryotic glycosites prediction

Home Galaxy Platform Shared Data Help

Overview

GlycoPP 2.0 is a highly accurate galaxy-based platform made available for the prediction of glycosylation sites in prokaryotic proteins. The GlycoPP 2.0 models are developed utilising [ProGlycProt V2.0](http://www.proglycprot.org/) database which is a repository of manually curated and experimentally characterized prokaryotic glycoproteins and protein glycosyltransferases respectively.

GlycoPP 2.0 is an updated version of GlycoPP (<http://crdd.osdd.net/raghava/glycopp/>)

N-linked Glycosylation O-linked Glycosylation

The service is free to use and does not require any registration/license for the analysis

Developed & Hosted by: CSIR-Institute of Microbial Technology, Sector-39A, Chandigarh, India (160036), <https://www.imtech.res.in>

Figure: 1 GlycoPP 2.0 home page.

2.2 GlycoPP 2.0 Galaxy landing page- The user can access the landing page of the GlycoPP 2.0 Galaxy interface by following the link- http://datascience.imtech.res.in/alkarao/glycopp_v2/. This page allows the user to perform predictions for N & O linked glycosylation employing developed models. The galaxy landing page of the GlycoPP 2.0 is shown in Figure 2 below.

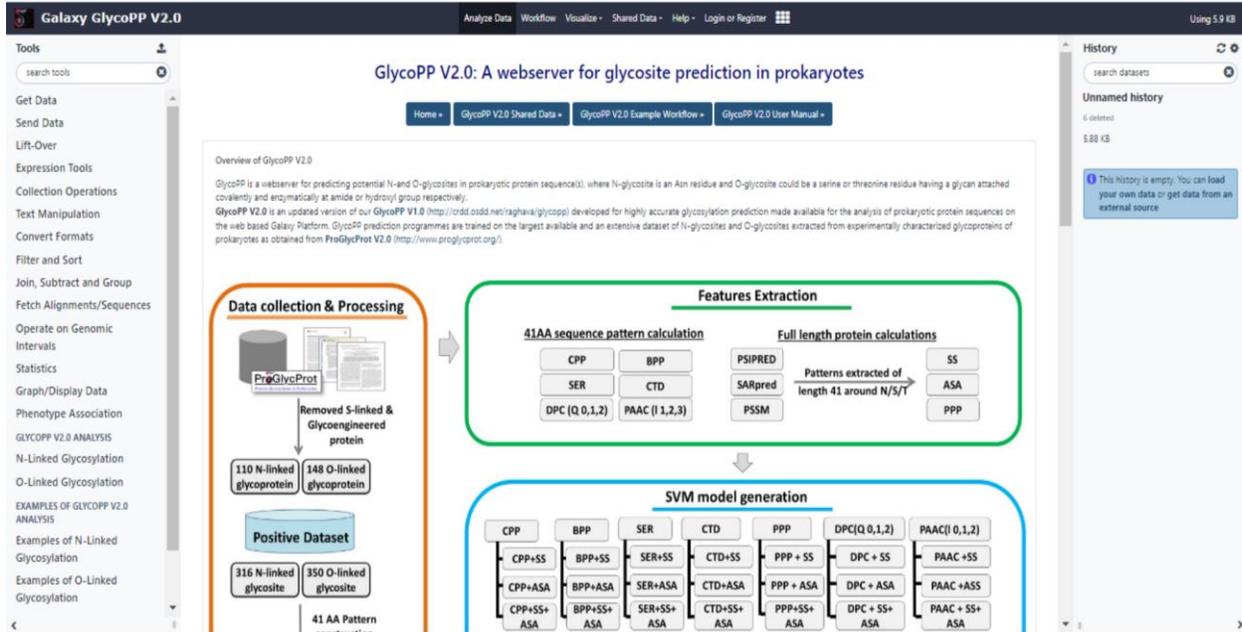


Figure: 2 GlycoPP 2.0 Galaxy landing page.

User friendliness: Keeping in mind that some of the users may not be familiar with Galaxy, we have created customized links directly to specific use cases from the main landing page to the Galaxy interface. This allows the user to navigate the Galaxy workflow system intuitively.

Note: The GlycoPP 2.0 web server is free to use and users can perform the predictions without any registration/license. In addition, we recommend the user to take an interactive tour of the Galaxy User-interface located at the bottom of the GlycoPP 2.0 Galaxy landing page as shown in Figure 3 to get familiar with the Galaxy platform. This tour provides a general overview of features in Galaxy.

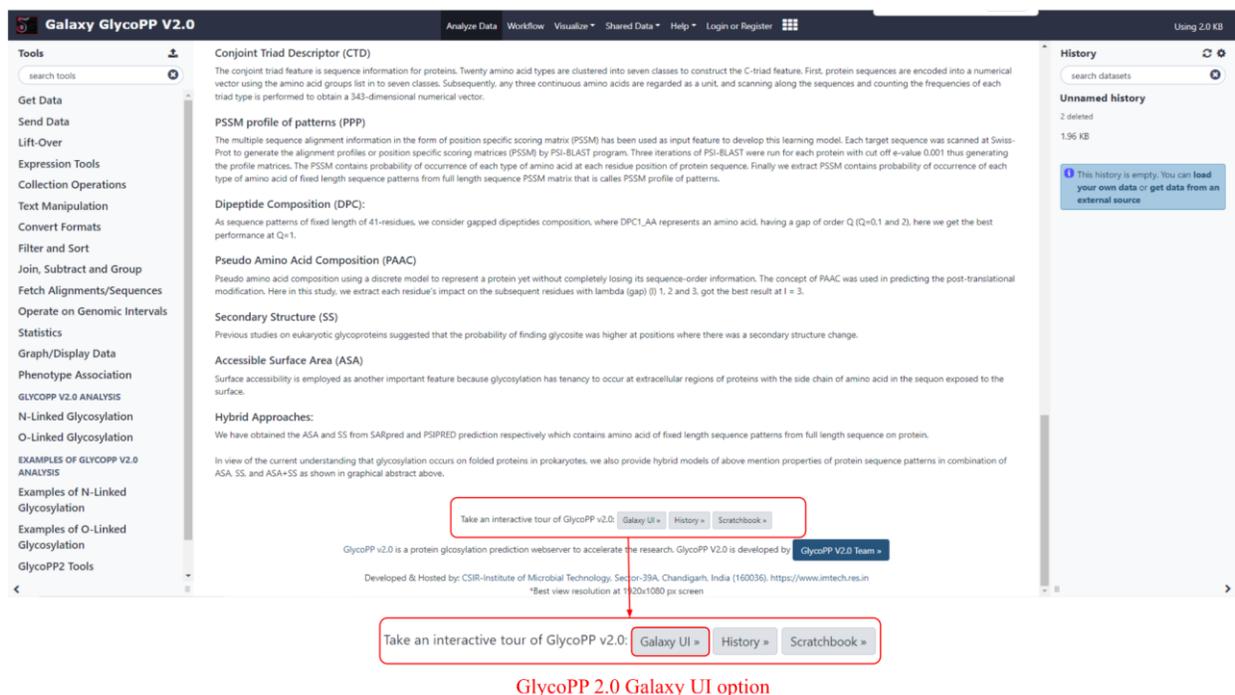


Figure: 3 Galaxy UI option at GlycoPP 2.0 Galaxy landing page.

3.0 Models Description

This discussion will allow the user to get an overview of the various features employed for the development of GlycoPP 2.0 SVM models for N-linked and O-linked glycosylation sites prediction in prokaryotic proteins.

1. **Composition Profile of Patterns (CPP):** Composition profile of patterns is the percentage frequencies of each amino acid in a fixed length sequence pattern. In this study we have generated motifs of length 41 amino acids in a way by which the glycosylated amino acid is placed at the middle of the motif i.e. 21st position.
2. **BPP Binary Profile Pattern (BPP):** In this approach, motifs of 41 amino acids were converted into binary form. Each type of the amino acid in motif was represented by a vector of dimension 20 (e.g. Ala by 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0; Cys by 0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0).
3. **Shannon Entropy of Residues (SER):** In order to understand the structural orchestration of sequences i.e., propensity towards order and disorder, the Shannon entropy (SE) score was calculated for each consensus sequence. It is

evident that entropy possesses an idea of the disorder and is directly proportional to the rate of disorder i.e., if the disorder increases, it signifies higher entropy.

- 4. Conjoint Triad Descriptor (CTD):** The conjoint triad descriptor is a kind of sequence information for proteins in which the twenty amino acids are clustered into seven classes based on their physicochemical properties to construct the C-triad feature. Briefly the protein sequences are encoded into a numerical vector using the seven classes of amino acids. Subsequently, any three continuous amino acids are regarded as a unit, and scanning along the sequences and counting the frequencies of each triad type is performed to obtain a 343-dimensional numerical vector.
- 5. PSSM Profile of Patterns (PPP):** The multiple sequence alignment information in the form of position specific scoring matrix (PSSM) has been used as an input feature to develop this learning model. Each target sequence was scanned at Swiss-Prot to generate the alignment profiles or position specific scoring matrices (PSSM) by PSI-BLAST program. Three iterations of PSI-BLAST were run for each protein with e-value cut-off 0.001, thus generating the profile matrices. The PSSM provides the probability of occurrence of each type of amino acid at each residue position in a protein sequence. Finally, we extracted the PSSM containing the probability of occurrence of each type of amino acid for fixed length sequence patterns from the full-length sequence PSSM matrix.
- 6. Dipeptide Composition (DPC):** As sequence patterns of fixed length of 41-residues, we considered gapped dipeptides composition, where DPC1_AA represents an amino acid, having a gap of order Q (Q=0,1 and 2), here we got the best performance at Q=1.
- 7. Pseudo Amino Acid Composition (PAAC):** Pseudo amino acid composition uses a discrete model to represent a protein without completely losing its sequence-order information. The concept of PAAC has been used in predicting the post-translational modification and in this study, we extracted each residue's impact on the subsequent residues with lambda (gap) (l) 1, 2 and 3 and got the best result at l = 3.

- 8. Secondary Structure (SS):** Previous studies on eukaryotic glycoproteins suggested that the probability of finding glycosite was higher at positions where there was a secondary structure change thus, we took this feature into account to predict the glycosylation sites.
- 9. Accessible Surface Area (ASA):** Surface accessibility of amino acids is regarded as an important feature as glycosylation has tendency to occur at extracellular regions of proteins with the side chain of amino acid in the sequon exposed to the surface.
- 10. Hybrid Models:** We have utilized the hybrid approaches in this study which involves using more than one feature to perform the prediction and assess their impact in improving overall accuracy of the developed model.

4.0 N-linked & O-linked glycosylation site prediction models of GlycoPP 2.0

The user can perform N-linked glycosylation sites prediction employing four different models namely **BPP** (Binary Profile Pattern), **BPP+ASA** (Binary Profile Pattern + Accessible Surface Area), **BPP+SS** (Binary Profile Pattern + Secondary Structure) and **BPP+ASA+SS** (Binary Profile Pattern + Accessible Surface Area + Secondary Structure). On the other hand, the user can predict the O-linked glycosylation sites employing six different types of models namely **CTD** (Conjoint Triad Descriptor), **PAAC** (Pseudo Amino Acid Composition based prediction), **SER** (Shannon Entropy of Residues), **CPP+SS** (Composition Profile of Patterns + Secondary Structure) based prediction, **DPC+SS** (Dipeptide Composition + Secondary Structure) and **DPC+ASA** (Dipeptide Composition Accessible Surface Area).

In the following tutorials we have considered the example of N-linked glycosylation sites prediction using **BPP** (Binary Profile Pattern) model for reference.

5.0 Running GlycoPP 2.0 with example files

The user can refer to the examples of the N-linked and O-linked glycosylation sites prediction by exploring the *Examples of GlycoPP 2.0 Analysis* option under the *Tools* panel. The work flow to perform the prediction and analysis by using the example files is summarized below:

1. From the *Tools* panel of the **GlycoPP 2.0 Galaxy landing page** the user can choose either *Examples of N-linked Glycosylation* or *Examples of O-linked Glycosylation* option. For this tutorial *Examples of N-linked Glycosylation* was taken as reference as shown in Figure 4.

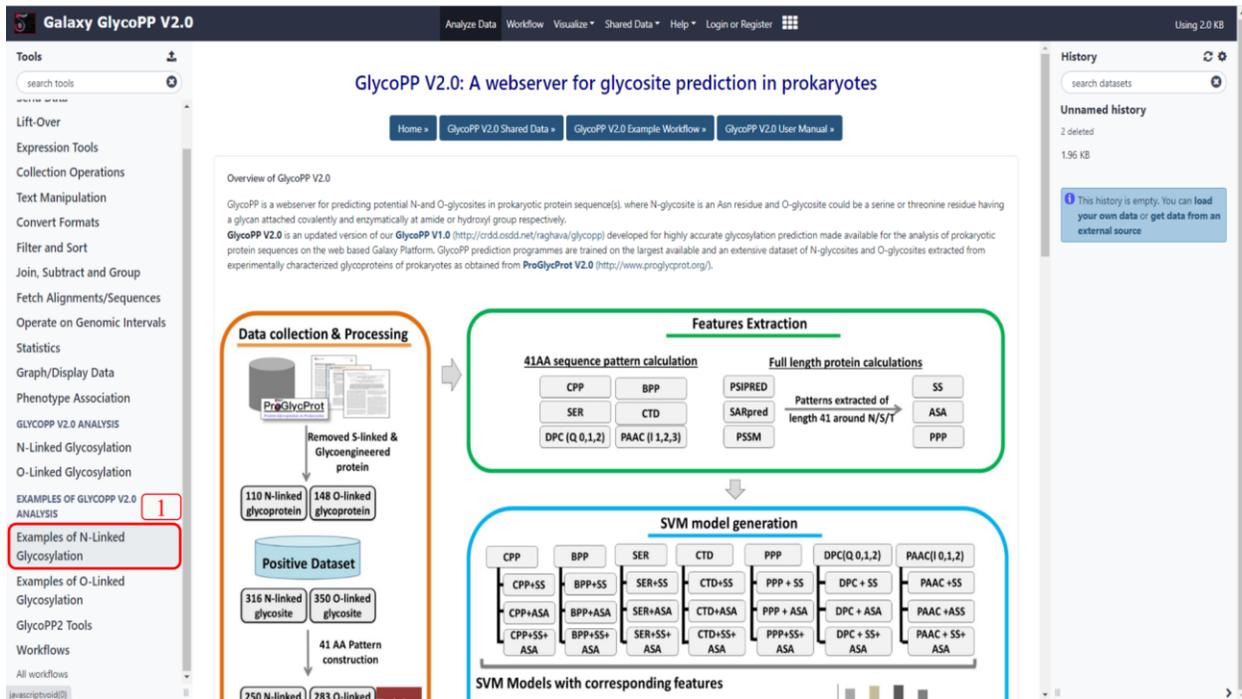


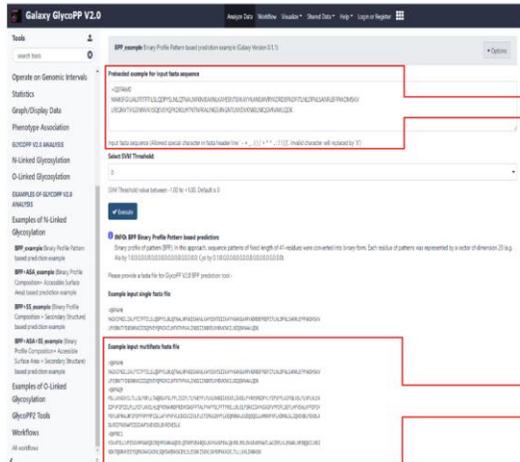
Figure: 4 Examples of N-linked Glycosylation option at the tools panel of GlycoPP 2.0 Galaxy landing page.

2. Upon selecting the *Examples of N-linked Glycosylation*, a drop-down menu will appear showing all of the models which can be used for the prediction. For this demonstration we are taking the *BPP (Binary Profile Pattern) based prediction example* model for N-linked glycosylation prediction into account as shown in Figure 5.

The screenshot displays the Galaxy GlycoPP V2.0 web interface. On the left, a sidebar lists various tools, with 'Examples of N-Linked Glycosylation' highlighted in a red box and labeled with a red '2'. The main workspace shows the configuration for the 'BPP_example Binary Profile Pattern based prediction example (Galaxy Version 0.1.1)' tool. It includes a 'Preloaded example for input fasta sequence' section with a text area containing a protein sequence. Below this is a 'Select SVM Threshold:' dropdown menu set to '0', with a note that the threshold value ranges from -1.00 to +1.00. An 'Execute' button is visible. Further down, there are sections for 'Example input single fasta file' and 'Example input multifasta fasta file', each with a text area containing protein sequences. The right sidebar shows a 'History' section with 'Unnamed history' and a message: 'This history is empty. You can load your own data or get data from an external source'.

Figure: 5 BPP (Binary Profile Pattern) based prediction example model for N-linked glycosylation prediction window.

- Once the user selects BPP (Binary Profile Pattern) model under *Examples of N-linked Glycosylation* the prediction execution window will appear comprising of *Preloaded example for input fasta sequence* (3.a), *Select SVM Threshold* (3.b), *Execute* (3.c) and representation of the input file and output file/prediction results along with other parameters (3.d) as shown in Figure 6.



Preloaded example for input fasta sequence

```
>QBPAM0
MAIKIFGILIALFTITFTILSLQDPYSYLNQTNALNFKNIEAKNLKAYESNTSIKAYYKANSWVRVADRDFNDFTLNLDFNLSANRLEFFNKDMSKV
LFEQNVITYGSMNVKISQEVYQPKDKILHTNTNFKALINGSINGNTLNYDVKNKLNQGVNIAWLQDK
```

Input fasta sequence (Allowed special character in fasta header line: ' - _ . () / + * ^ , ; ? ! []'. Invalid character will be replaced by 'X')

Preloaded example input window

Example input multifasta fasta file

```
>QBPAM0
MAIKIFGILIALFTITFTILSLQDPYSYLNQTNALNFKNIEAKNLKAYESNTSIKAYYKANSWVRVADRDFNDFTLNLDFNLSANRLEFFNKDMSKV
LFEQNVITYGSMNVKISQEVYQPKDKILHTNTNFKALINGSINGNTLNYDVKNKLNQGVNIAWLQDK
>QBPAA09
MSLKNISSYILTLLSLFGFLLTWQRSFSLFLLISIFLTFHEFFLFLKRNIIKEATLIGSLFYRVSMGDFLYLFFSFLAIFGLVSLFLNLFNLEK
IDFVFIIFLPLMLIFLKKELHLQFVDNAYNDFRIVIASFTALFYVGLFFTYNELLNLEFSRKTIAVKSASFVYDFLSEFLHFVSNLKFIFSY
FGYGLFRALNFIQDFNFFMFCSLLAFVNFVFLKTKIKIIVLFLCFIMVLGNVFLKEQRNALKSEQEQLLMMNINFLKDNINLSIQEKDLFEKDLK
DLREIFKKNIAFETIGIWFSEKEDLEKRINESLK
>QBPBC1
MIKFKLLVFTSSVVFAGQDCEQYFARKAQTELQTRFDEARQSL EAYKASFEALQKERLENLEKKEAEVWATLAKTEELKENARLVEEQKILNLSI
NDKTQGRVKEIYSQMKDAIADVLSQMDAEDASKIMLSLESRKISGVLKMDPKKASELTL LKINLDNINASHI
```

Example input multiple sequence fasta file

Figure: 6 Preloaded example input window and example multiple input fasta file.

5. The next option after this, *Select SVM Threshold* comes with a drop-down menu with values ranging from -1.0 to 1.0. The default threshold of 0 indicates that the *Residue score/prediction probability* value greater than 0 will designate that residue as *Potential Glycosylated* while a score less than 0 will allow the model to assign that residue as *Non-glycosylated*. Conclusively the SVM threshold value acts as a cut-off point to perform the binary classification and the user can try different threshold values as per their requirement by prioritizing false positives or false negatives. The select SVM threshold window is shown in Figure 7.

The screenshot displays the Galaxy GlycoPP V2.0 web interface. The main content area is titled "BPP+ASA_example (Binary Profile Composition+ Accessible Surface Area) based prediction example (Galaxy Version 0.1.1)". It features a "Preloaded example for input fasta sequence" text area containing a protein sequence:


```
>Q0PAM0
MAIKIFGLIALFTITFTILSLQDPYSLNLQTNALNFKNIEAKNLKAYESNTSIIKAYYKANSWVRYADRFDFITL
NLDFNLSANRLEFFNKDMSKV
LFEGNVTYIGSNVVKIISQVEYQPKDKILHTNTNFKALINGSIINGNTLNVDVKNKILNIQGVNAWLQDK
```

 Below the text area is a label "Input fasta sequence (Allowed special character in fasta header line '- = _ . () / + * ^ , ; : ! []'. Invalid character will be replaced by 'X')". A red box highlights the "Select SVM Threshold:" dropdown menu, which is currently set to "0". A red circle with the number "5" is placed to the right of this dropdown. Below the dropdown, a note states "SVM Threshold value between -1.00 to +1.00. Default is 0". An "Execute" button is visible below the threshold selection. An information box at the bottom provides details about the BPP+ASA method. The left sidebar contains navigation options like "Tools", "Operate on Genomic Intervals", and "Statistics". The right sidebar shows a "History" section with a message: "This history is empty. You can load your own data or get data from an external source".

Figure: 7 Select SVM threshold window at the GlycoPP 2.0 interface.

6. Finally, the user can complete the submission process and initialize the prediction by clicking on the *Execute* (6.a) button which will redirect the user to a successful job submission window (6.b) from where the user can find their job queue status and the output file (6.c) being generated. The input execution and job submission window is shown in Figure 8 & 9.

the residues involved in glycosylation (N (Asn) for N-linked glycosylation and S (Ser)/T (Thr) for O-linked glycosylation). The second type of output (7.b) will also be in a tabular form which will show the information of the residues (Residue position, one letter amino acid code of the residue along with following two residues in respective protein sequence, Prediction score and prediction status) which have been designated as *Potential Glycosylated* by the developed model. The user will have the access to view the above discussed results on the interface itself and can also save the results in tsv format.

The output file for the current submission signifies that there were 24 N (Asn) residues in the protein sequence out of which the residues at the position 51, 84, 105, 141 have been predicted as Potential Glycosylated. The output result files for the above are shown in Figure 10, 11 and 12.

The screenshot displays the Galaxy GlycoPP V2.0 interface. The main panel shows the output of a job, including a header with the protein name >Q0PAM0 and its length (171). Below this, the potential N-linked glycosylation sites are listed as MAIKIFGILALFTITFTILSLQDPYSLNLQTNALNFKNIEAKNLKAYESNTSIKAYYKANSWVYRDRDEFNDFITLNLDFNLSANRLEFFNKDMSKV and LFGENVTYIGSNVVKIISQEVYQPKDKILHTNTNFKALINGSIINGNTLNYDVKNKILNIQGVNAWLQDK. The GlycoPP v2.0 Prediction Method is BPP and the SVM Threshold is 0. A table follows, listing the position, residue, score, and prediction for each residue. The table shows that residues at positions 51, 84, 105, and 141 are predicted as 'Potential Glycosylated', while all other residues are 'Non-glycosylated'. On the right side, the 'History' panel shows a list of jobs, with '4: BPP: Glycosylation' and '3: BPP_example' highlighted. A red box labeled '7.a' is drawn around the '3: BPP_example' entry in the history panel, with a red arrow pointing from it to the main output table.

Position	Residue	Score	Prediction
29	NLQ	-0.68692359	Non-glycosylated
33	NAL	-0.72174881	Non-glycosylated
36	NFK	-0.12207736	Non-glycosylated
39	NIE	-0.32470975	Non-glycosylated
44	NLK	-0.20398314	Non-glycosylated
51	NTS	1.0002212	Potential Glycosylated
62	NSW	-0.52280092	Non-glycosylated
74	NDF	-0.35368478	Non-glycosylated
80	NLD	-0.8578013	Non-glycosylated
84	NLS	0.71732003	Potential Glycosylated
88	NRL	-0.36008792	Non-glycosylated
94	NKD	-0.15063364	Non-glycosylated
105	NVT	0.63372566	Potential Glycosylated
112	NNV	-0.70692049	Non-glycosylated
113	NVK	-0.50830369	Non-glycosylated
133	NTN	-0.17287851	Non-glycosylated
135	NFK	-0.53119115	Non-glycosylated
141	NGS	0.73101075	Potential Glycosylated
146	NGN	-0.46322435	Non-glycosylated

Figure: 10 Output file of the submitted job showing complete run information and tabulated prediction result including position, residue, score and prediction result.

Galaxy GlycoPP V2.0

>QOPAM0-Position	>QOPAM0-Residue	BPP-T0-Score	BPP-T0-Prediction
>QOPAM0-Position	>QOPAM0-Residue	BPP-T0-Score	BPP-T0-Prediction
51	NTS	1.0002212	Potential Glycosylated
84	NLS	0.71732003	Potential Glycosylated
105	NVT	0.63372566	Potential Glycosylated
141	NGS	0.73101075	Potential Glycosylated

History

Unnamed history

2 shown, 2 deleted

3.92 KB **7.b**

4: BPP: Glycosylation

4 lines, 1 comments

format: tsv, database: ?

```

1.:>QOPAM0-Position 2.:>QOPAM0-Residue 3.BPP-T0-
>QOPAM0-Position >QOPAM0-Residue BPP-T0-S
51 NTS 1.000221
84 NLS 0.717320
105 NVT 0.633725
141 NGS 0.731010

```

3: BPP_example

Figure: 11 Output file of the submitted job showing information of the residues which have been predicted as Potential glycosylated.

Galaxy GlycoPP V2.0

>QOPAM0-Position	>QOPAM0-Residue	BPP-T0-Score	BPP-T0-Prediction
>QOPAM0-Position	>QOPAM0-Residue	BPP-T0-Score	BPP-T0-Prediction
51	NTS	1.0002212	Potential Glycosylated
84	NLS	0.71732003	Potential Glycosylated
105	NVT	0.63372566	Potential Glycosylated
141	NGS	0.73101075	Potential Glycosylated

History

Unnamed history

2 shown, 2 deleted

3.92 KB **7.c**

4: BPP: Glycosylation

4 lines, 1 comments

format: tsv, database: ?

Download, view details and Run this job again options

```

1.:>QOPAM0-Position 2.:>QOPAM0-Residue 3.BPP
>QOPAM0-Position >QOPAM0-Residue BPP-T0-
51 NTS 1.0002
84 NLS 0.7173
105 NVT 0.6337
141 NGS 0.7310

```

3: BPP_example

33 lines, 1 comments

format: tsv, database: ?

```

1.:>QOPAM0
>QOPAM0
Potential N-Linked Glycosylated Sites:
PAIKKFGTLLALFTTFITLSDQPSYLNQTHALMPEZKAP
LFEQWTVYIGSNNKIIISQEVQPKDKLHHTHTFKALINGSL

```

Figure: 12 Download, view details and run this job again option for the generated output.

6.0 Running GlycoPP 2.0 with user supplied dataset

6.1 Data upload for prediction of glycosylation sites

The user can submit the query protein sequences by using the *upload file* from your computer option in the *Get Data* dropdown menu of the Galaxy Tools panel. Upon accessing the *Upload file* option the *Download from web or upload from disk* (6.1.a) pop-up will appear where the user can submit their data in various ways including *choose local file* (6.1.b) option. After selecting the dataset, the user can click on the *start* (6.1.c) to complete the data upload process. The screenshots for the above are shown in Figure 13, 14 and 15.

Galaxy GlycoPP V2.0

Analyze Data | Workflow | Visualize | Shared Data | Help | Login or Register

Tools

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- EBI SRA ENA SRA
- modENCODE fly server
- InterMine server
- Flymine server
- modENCODE modMine server
- MouseMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- ZebrafishMine server
- EuPathDB server
- HbVar Human Hemoglobin Variants and Thalassemias
- Send Data
- Lift-Over
- Enrichment Tools

GlycoPP V2.0: A webserver for glycosite prediction in prokaryotes

Home | GlycoPP V2.0 Shared Data | GlycoPP V2.0 Example Workflow | GlycoPP V2.0 User Manual

Overview of GlycoPP V2.0

GlycoPP is a webserver for predicting potential N- and O-glycosites in prokaryotic protein sequence(s), where N-glycosite is an Asn residue and O-glycosite could be a serine or threonine residue having a glycan attached covalently and enzymatically at amide or hydroxyl group respectively.

GlycoPP V2.0 is an updated version of our GlycoPP V1.0 (<http://crdd.osdd.net/raghava/glycopp>) developed for highly accurate glycosylation prediction made available for the analysis of prokaryotic protein sequences on the web-based Galaxy Platform. GlycoPP prediction programmes are trained on the largest available and an extensive dataset of N-glycosites and O-glycosites extracted from experimentally characterized glycoproteins of prokaryotes as obtained from ProGlycProt V2.0 (<http://www.proglycprot.org/>).

Data collection & Processing

ProGlycProt

Removed S-linked & Glycoengineered protein

110 N-linked glycoprotein | 148 O-linked glycoprotein

Positive Dataset

116 N-linked glycosite | 150 O-linked glycosite

Features Extraction

41AA sequence pattern calculation

CPP	BPP	PSIPRED	SS
SER	CTD	SARpred	ASA
DPC (Q 0,1,2)	PAAC (I 1,2,3)	PSSM	PPP

Full length protein calculations

Patterns extracted of length 41 around N/S/T

SVM model generation

CPP	BPP	SER	CTD	PPP	DPC(Q 0,1,2)	PAAC(I 0,1,2)
CPP+SS	BPP+SS	SER+SS	CTD+SS	PPP+SS	DPC+SS	PAAC+SS
CPP+ASA	BPP+ASA	SER+ASA	CTD+ASA	PPP+ASA	DPC+ASA	PAAC+ASA

Figure: 13 Upload file from computer option in the Get Data tools panel.

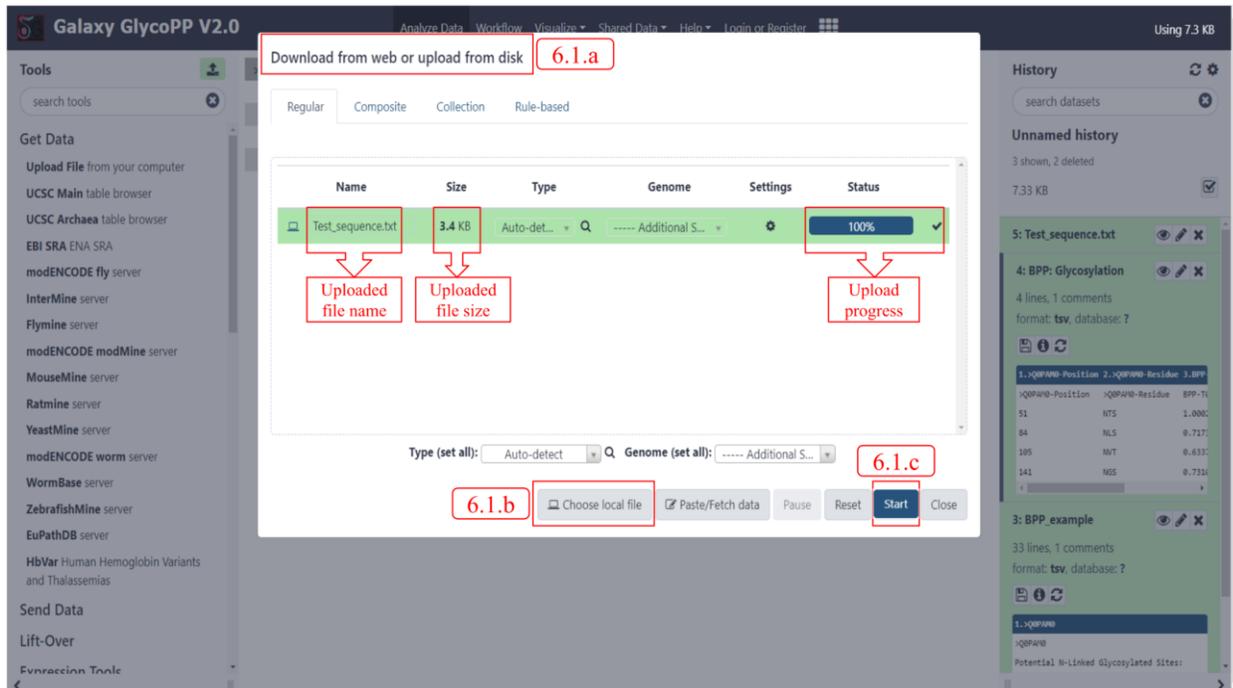


Figure: 14 Data upload from disk window when the user selects upload file from computer option.

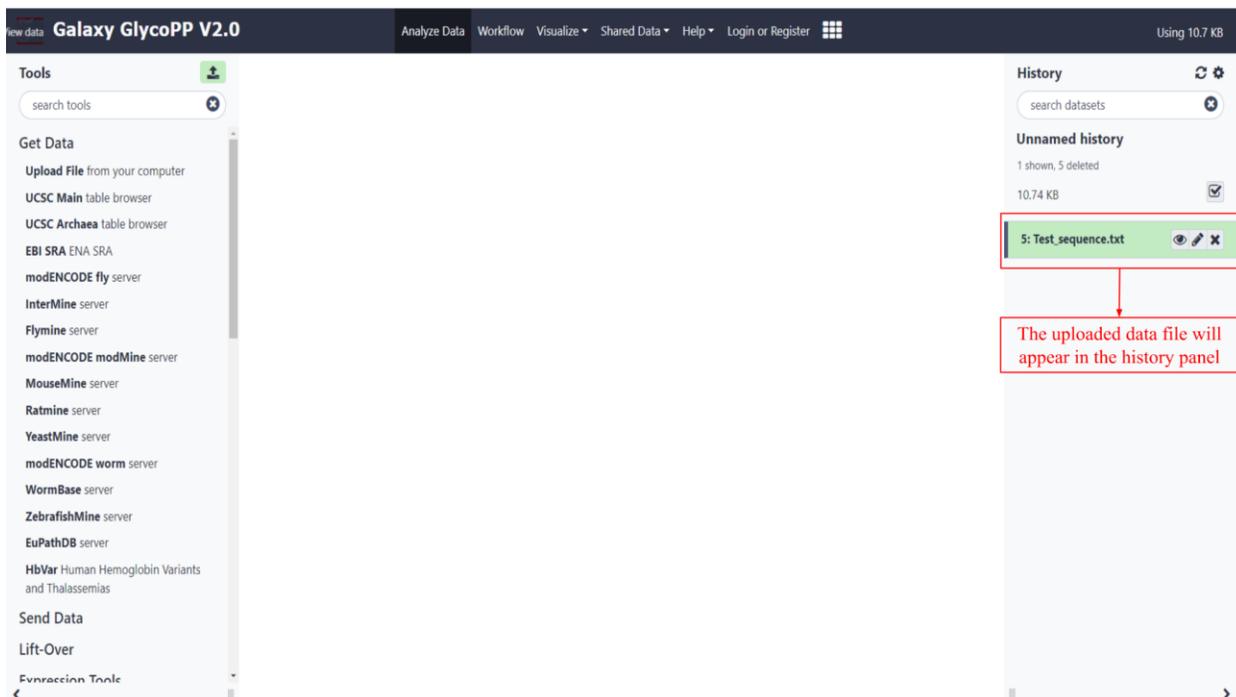


Figure: 15 The user can view, edit or delete the uploaded data file from the history panel.

Note: The user is required to submit the protein sequence in fasta format only

By using the *GlycoPP 2.0 Analysis* option under the *Tools* panel, the user can proceed to perform the N-linked and O-linked glycosylation sites prediction on the supplied dataset as described in the section 4.0 and 5.0 of the manual.

Example: N-linked glycosylation sites prediction using BPP (Binary Profile Pattern) model

1. The user needs to select the *BPP (Binary Profile Pattern) based prediction* option under *N-linked glycosylation* tab in *Tools* panel and select the uploaded data file from the drop-down menu under the *Input Fasta File* option as shown in Figure 7.

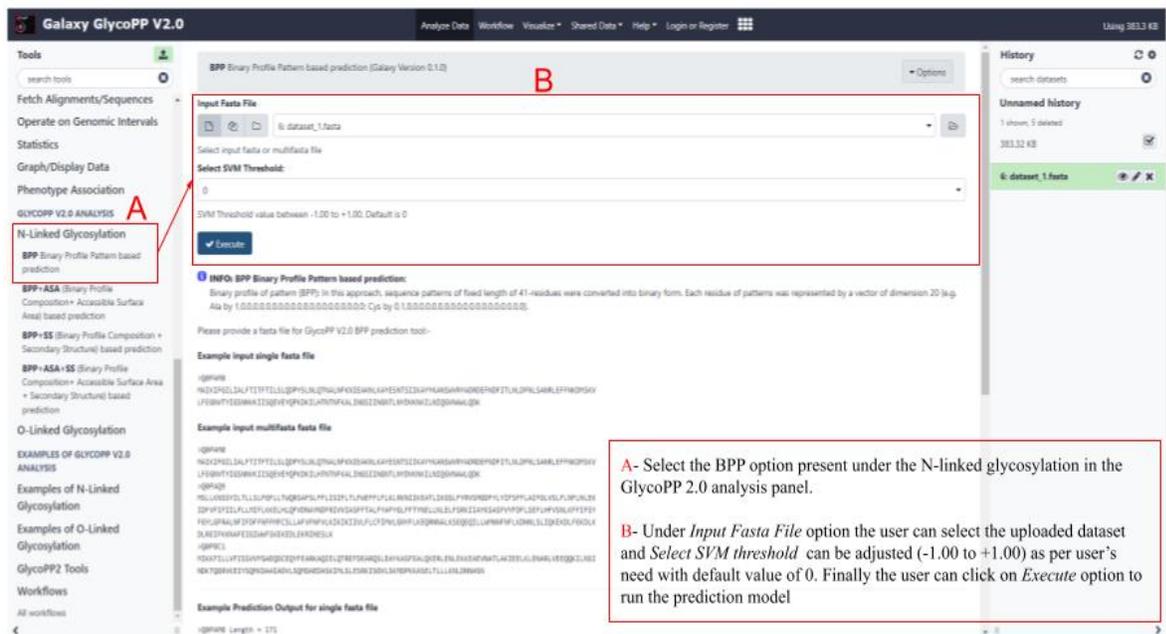


Figure 16: GlycoPP 2.0 model selection, Input file submission, SVM threshold selection and execution interface.

2. Upon execution of the job the user can track the status of the same from the history panel and upon completion download, view and re-run the job again. The output for the prediction is shown in Figure 8.

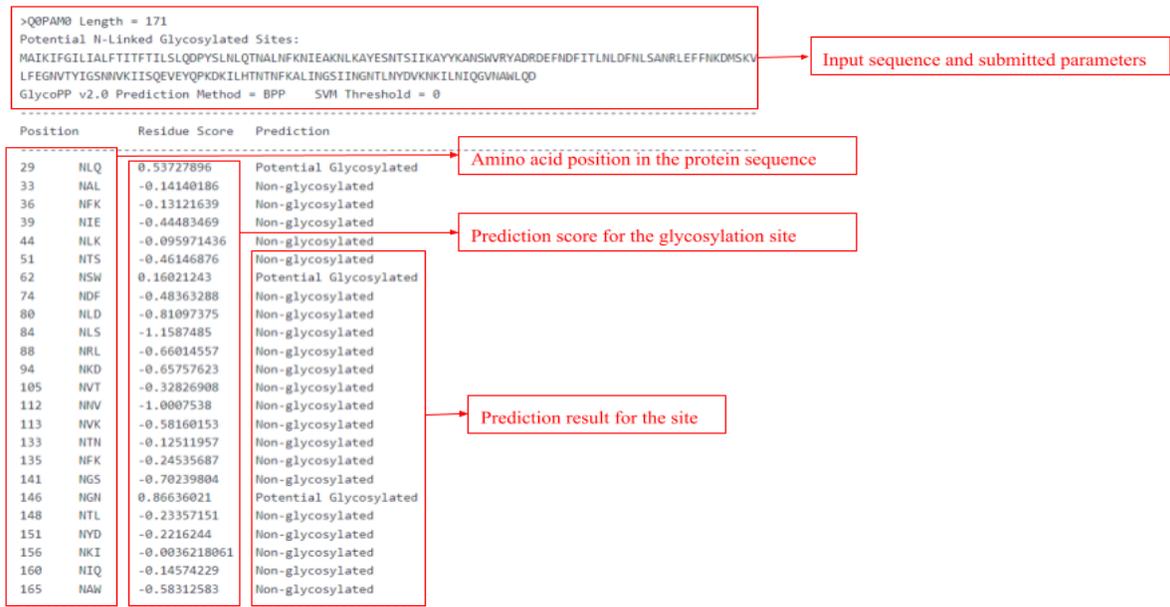


Figure 17: Screenshot of the GlycoPP 2.0 output page.

The user can select any combination of models to perform predictions. As GlycoPP2.0 also has tools for feature calculation, the users can generate models based on their dataset making the platform extremely scalable and not limited by availability of data at the time of generating models.

In case of any issues in following the tutorial or running GlycoPP2.0, you are welcome to contact raoalka@imtech.res.in or anshu@imtech.res.in



सीएसआईआर
CSIR
भारत का नवाचार इंजन
The Innovation Engine of India