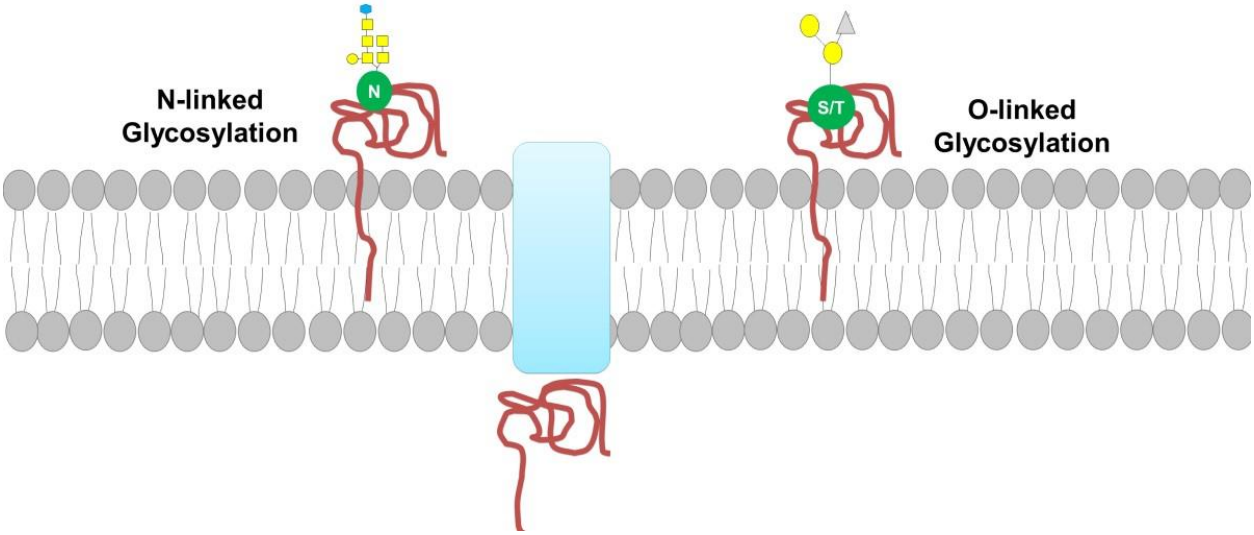
GlycoPP 2.0

User Manual Version 1.0

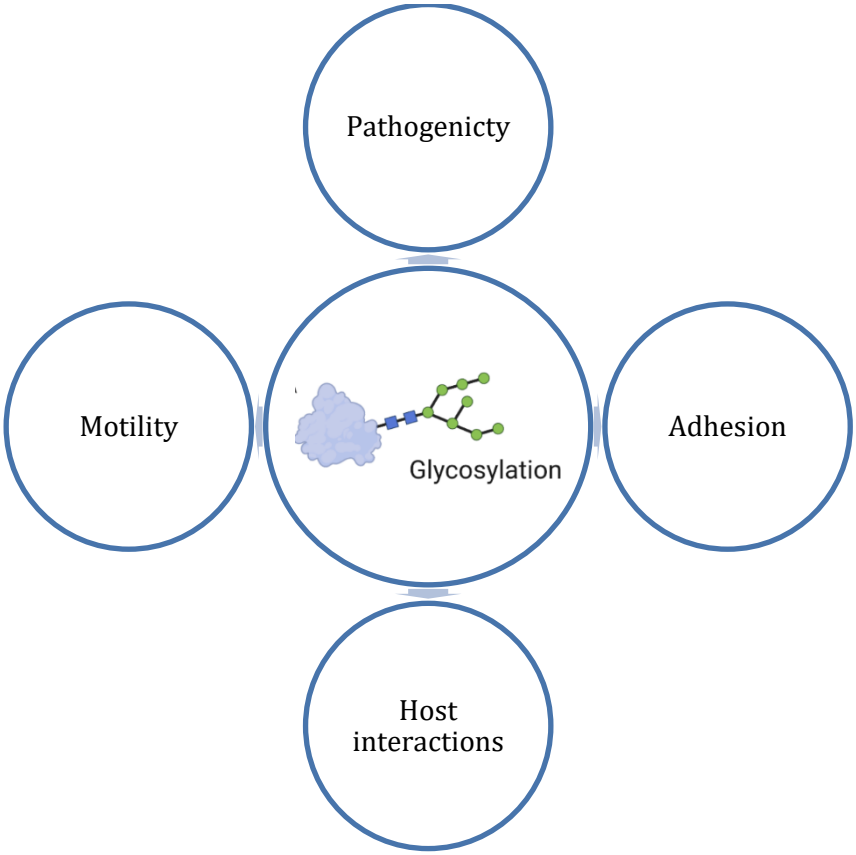


https://datascience.imtech.res.in/alkarao/glycopp_v2

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.

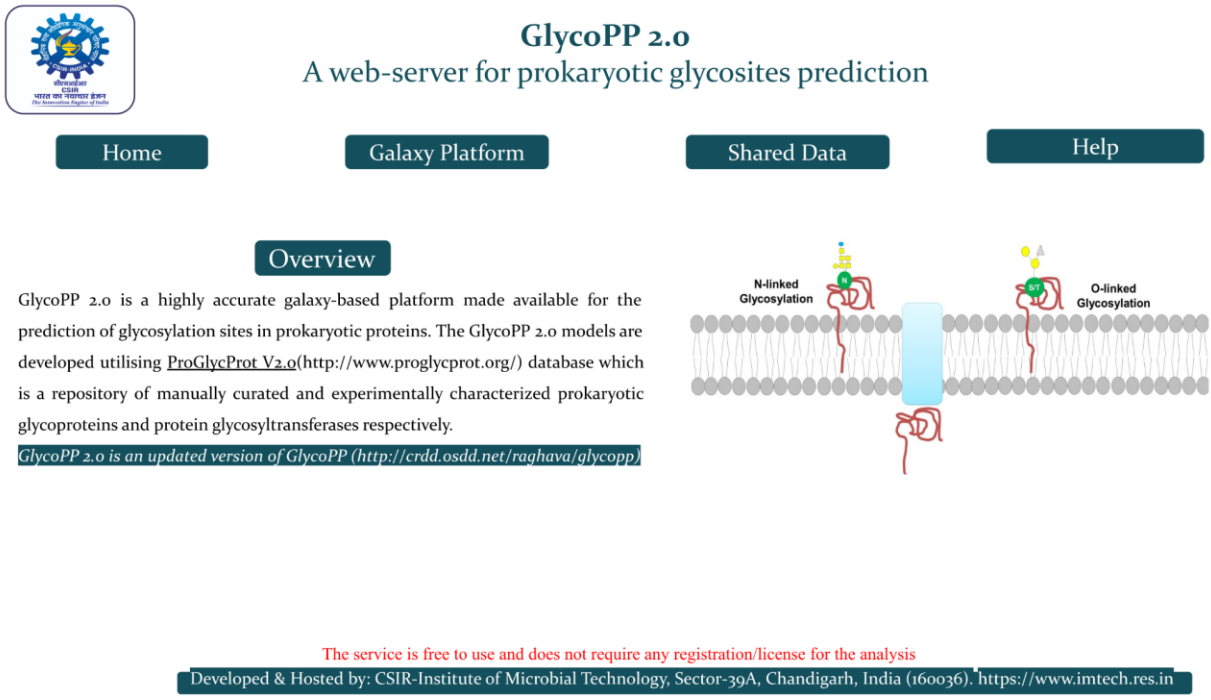


Figure: 1 GlycoPP 2.0 home page.

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.

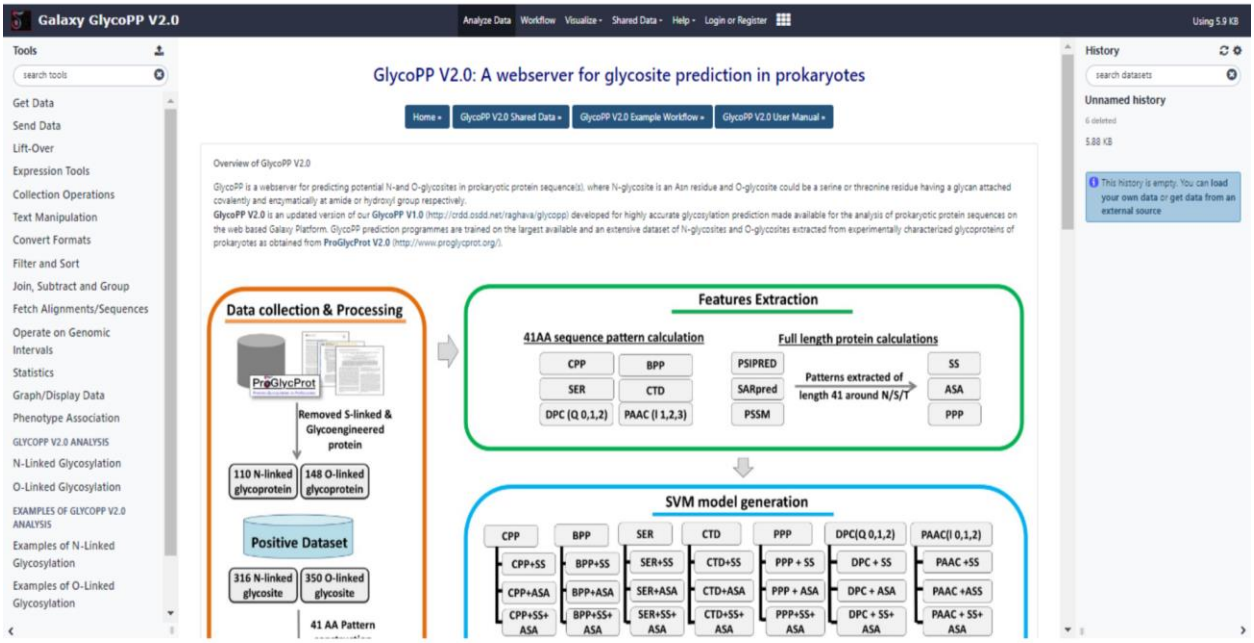
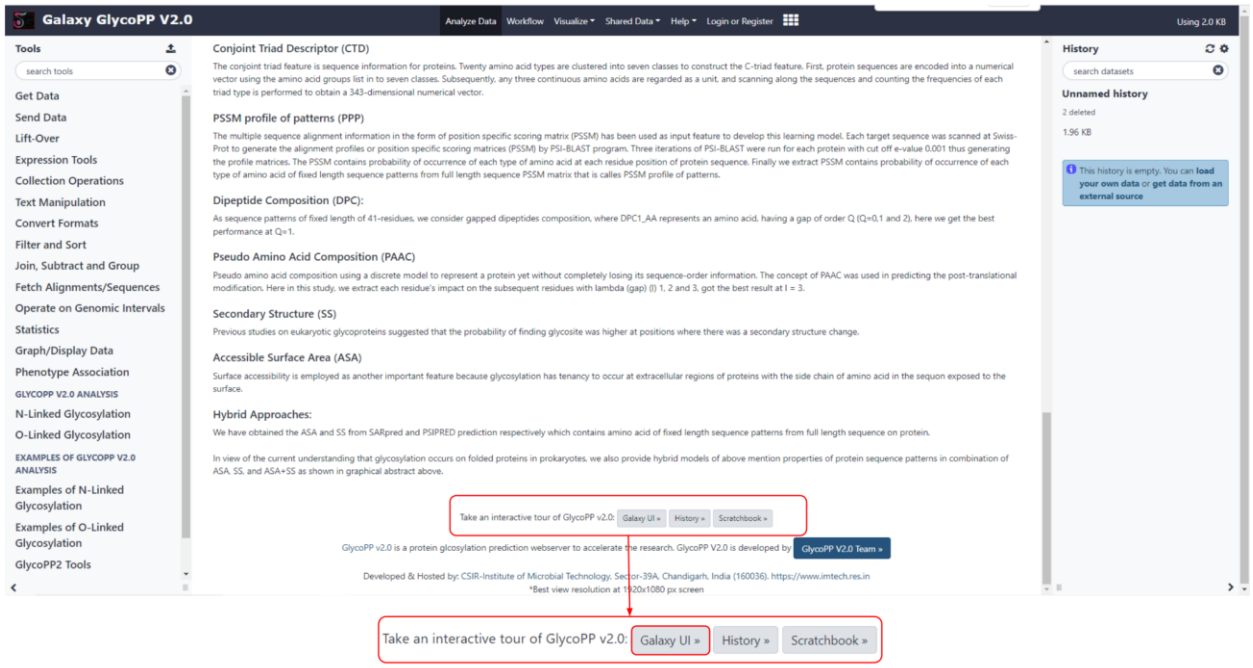


Figure: 2 GlycoPP 2.0 Galaxy landing page.

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.



GlycoPP 2.0 Galaxy UI option

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.

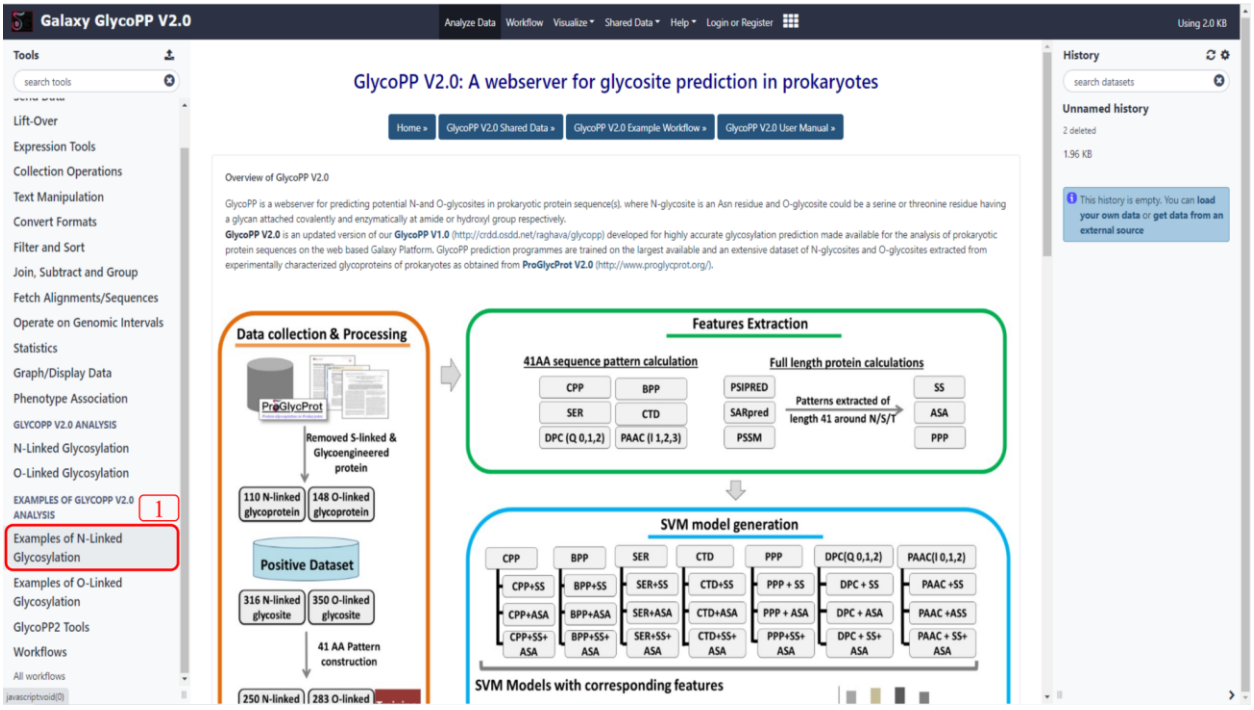


Figure: 4 Examples of N-linked Glycosylation option at the tools panel of GlycoPP 2.0 Galaxy landing page.

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.

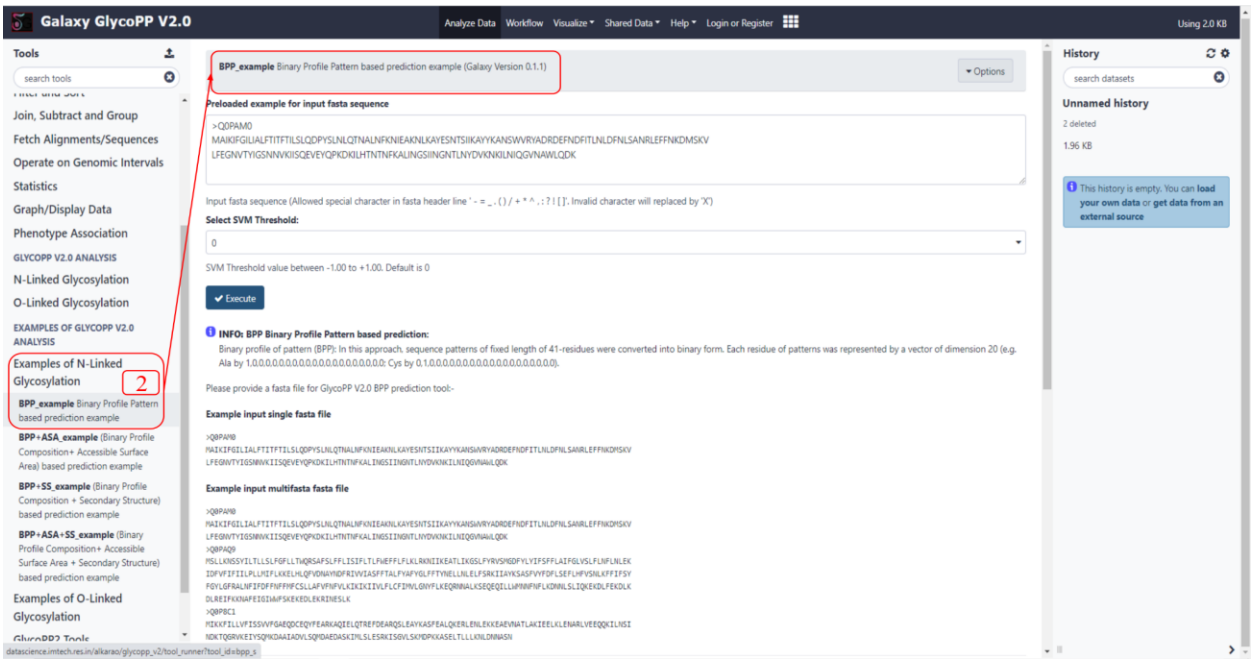


Figure: 5 BPP (Binary Profile Pattern) based prediction example model for N-linked glycosylation prediction window.

3. 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.

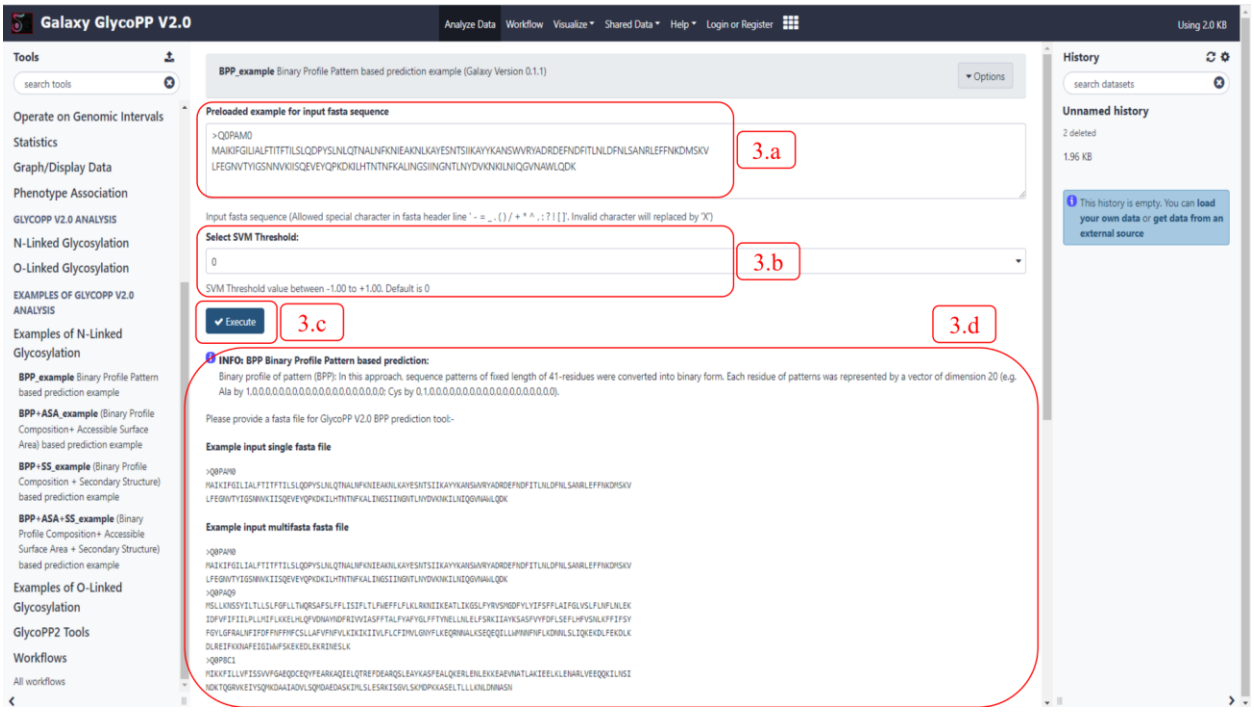


Figure: 6 GlycoPP 2.0 submission window comprising Preloaded example for input fasta sequence (3.a), Select SVM Threshold (3.b), Execute (3.c) and representation of the input file and output file/prediction results along with other parameters (3.d).

4. On the *Preloaded example for input fasta sequence* tab the user will see the preloaded input protein sequence in fasta format. The preloaded protein sequence having UniprotID- Q0PAM0 is a membrane protein of *Campylobacter jejuni* HB93-13 with 171 amino acids. The user can also use multiple fasta sequences by copying them from the *Example input multifasta fasta file* as shown in Figure 7.

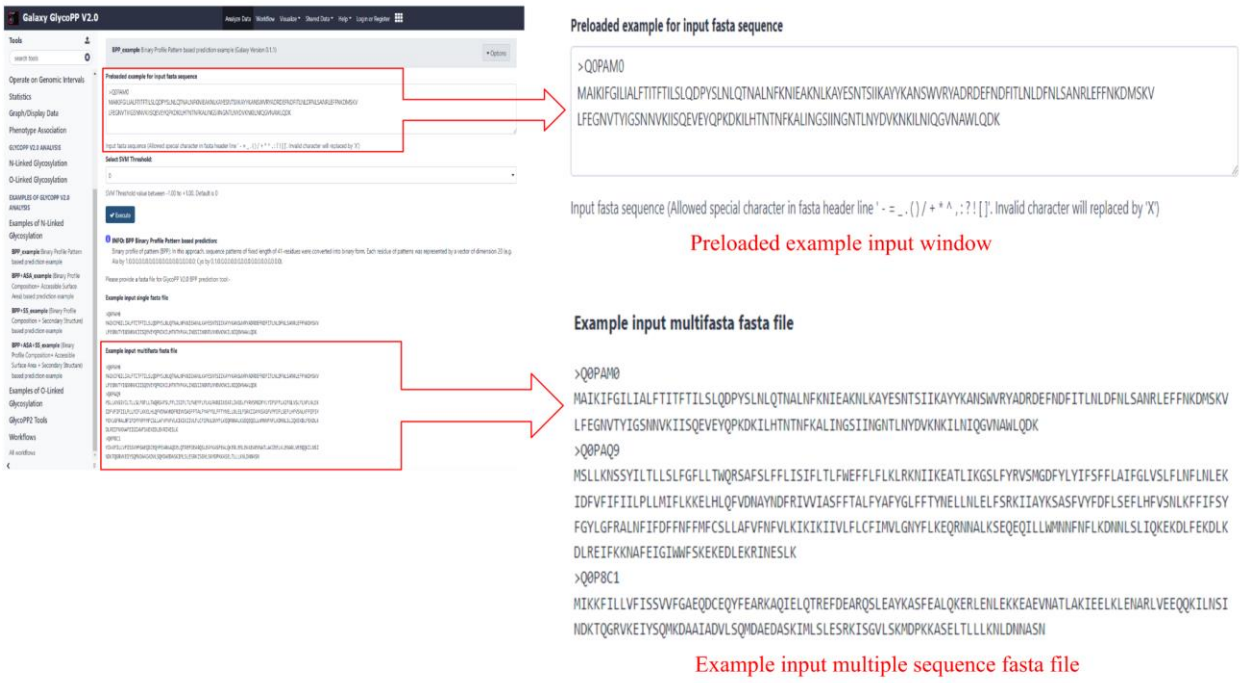


Figure: 6 Preloaded example input window and example multiple input fasta file.

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.

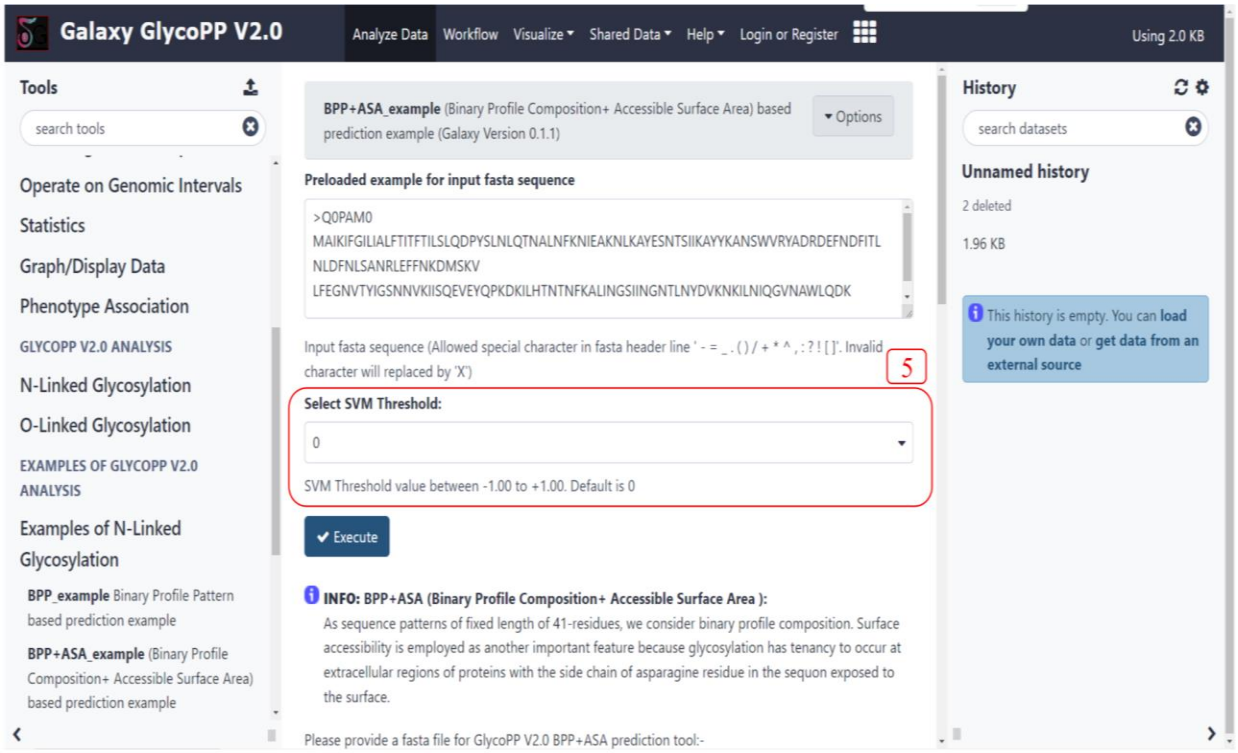


Figure: 7 Select SVM threshold window at the GlycoPP 2.0 interface.

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.

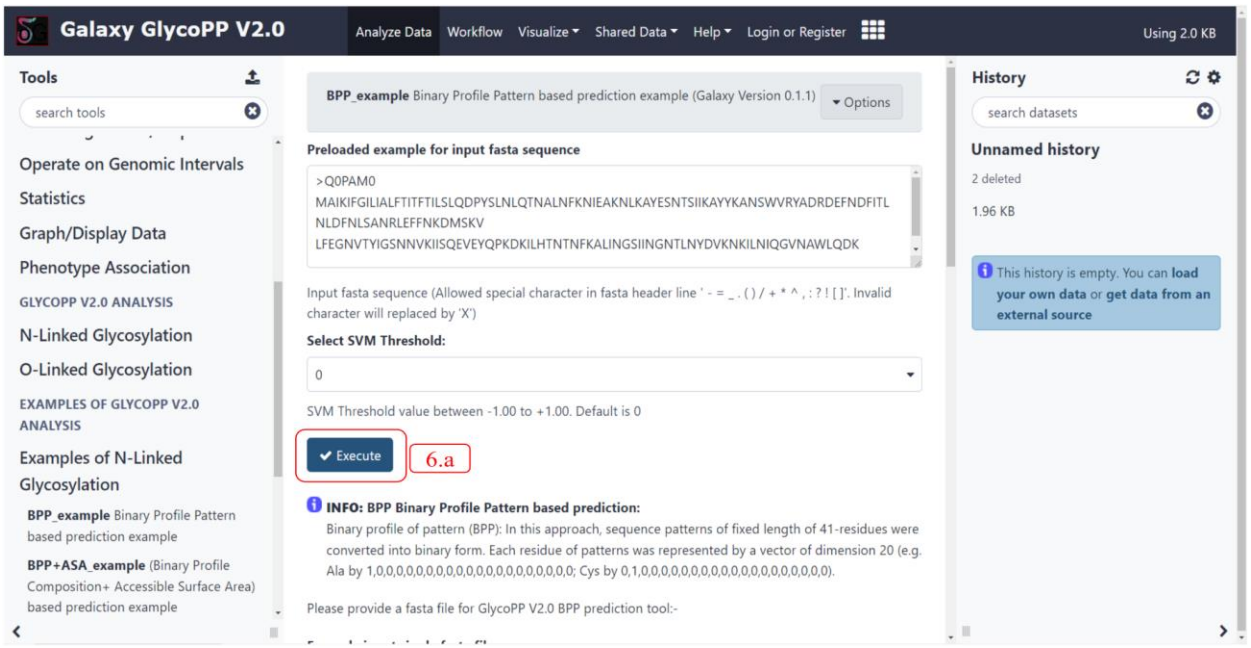


Figure: 8 Submission execution option of GlycoPP 2.0.

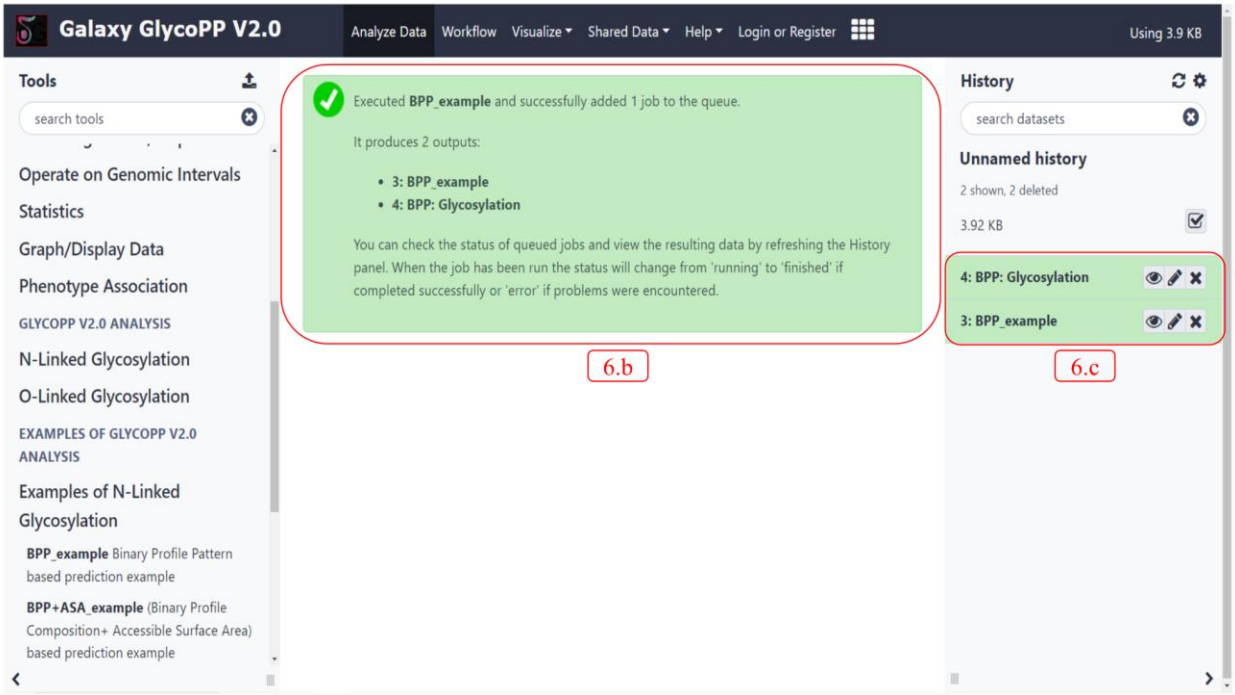


Figure: 9 Job submission status window of GlycoPP 2.0.

7. Once the submitted job status is completed the user will have the access to output from the history panel which includes *BPP_example* (7.a) showing complete run information including protein identifier, protein length, sequence, Type of prediction method used, SVM threshold and table showing predicted glycosylation status of

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.

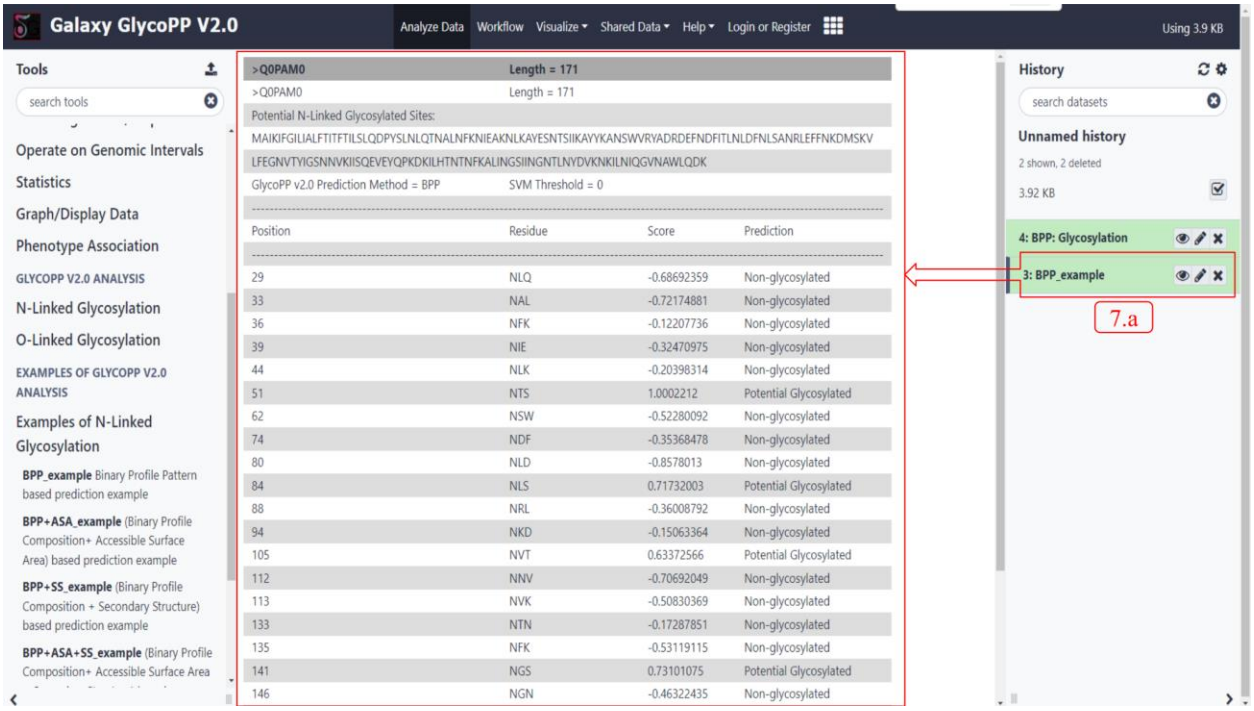


Figure: 10 Output file of the submitted job showing complete run information and tabulated prediction result including position, residue, score and prediction result.

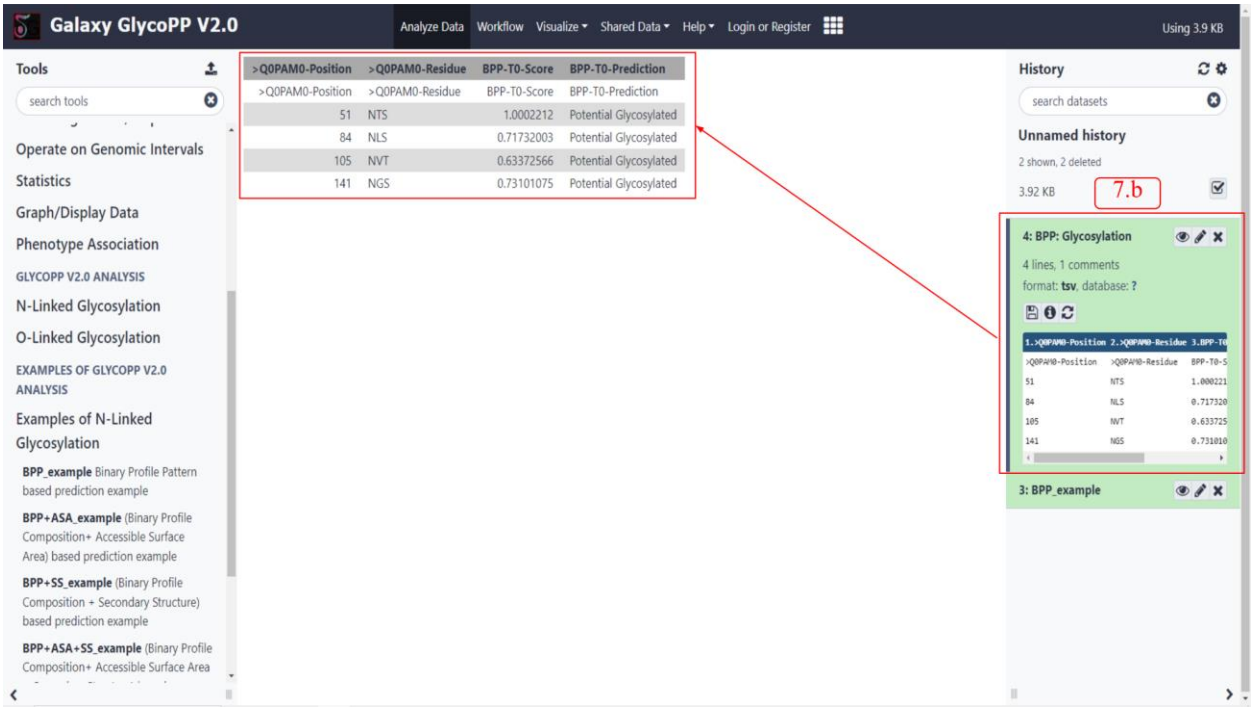


Figure: 11 Output file of the submitted job showing information of the residues which have been predicted as Potential glycosylated.

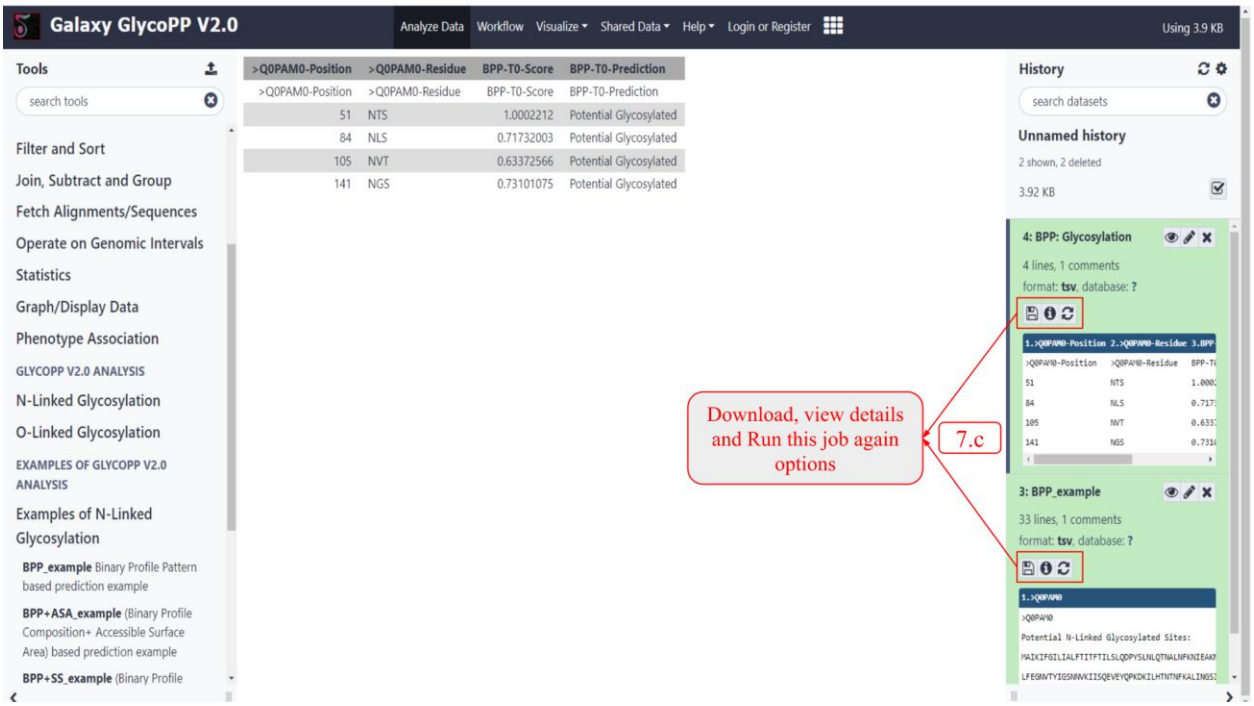


Figure: 12 Download, view details and run this job again option for the generated output.

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.

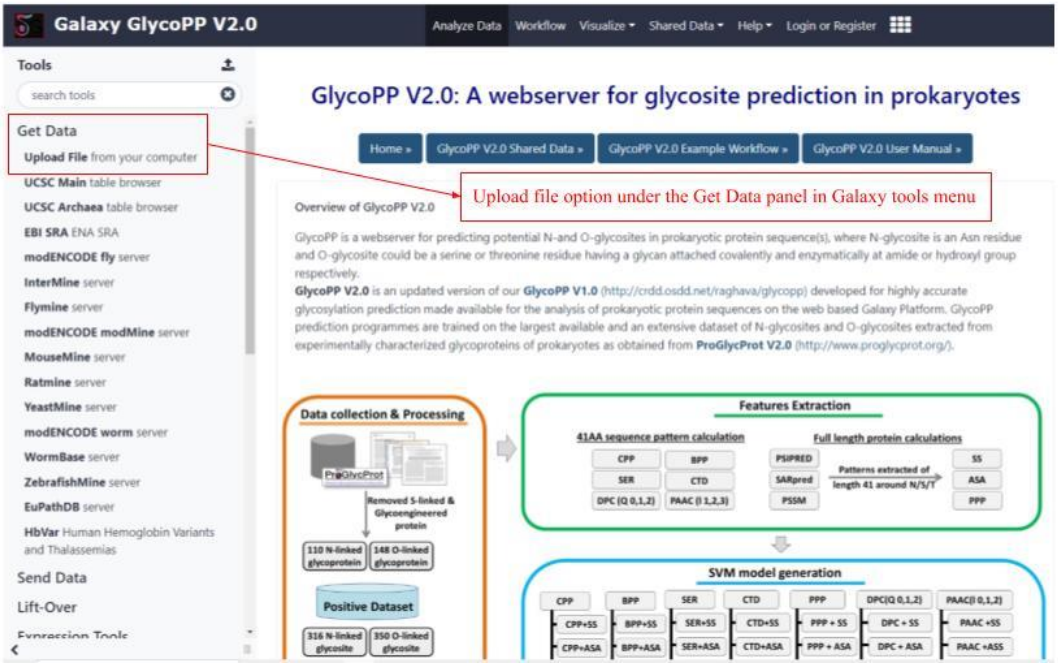


Figure: 13 Upload file from computer option in the Get Data tools panel.

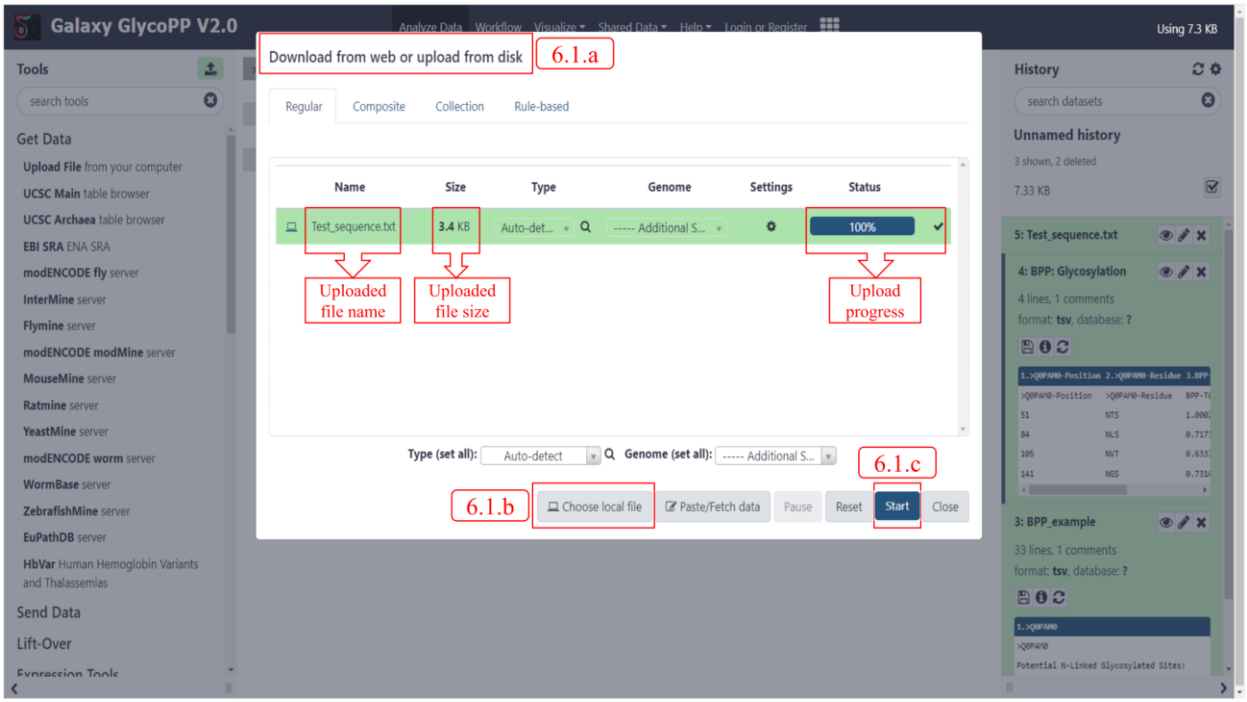


Figure: 14 Data upload from disk window when the user selects upload file from computer option.

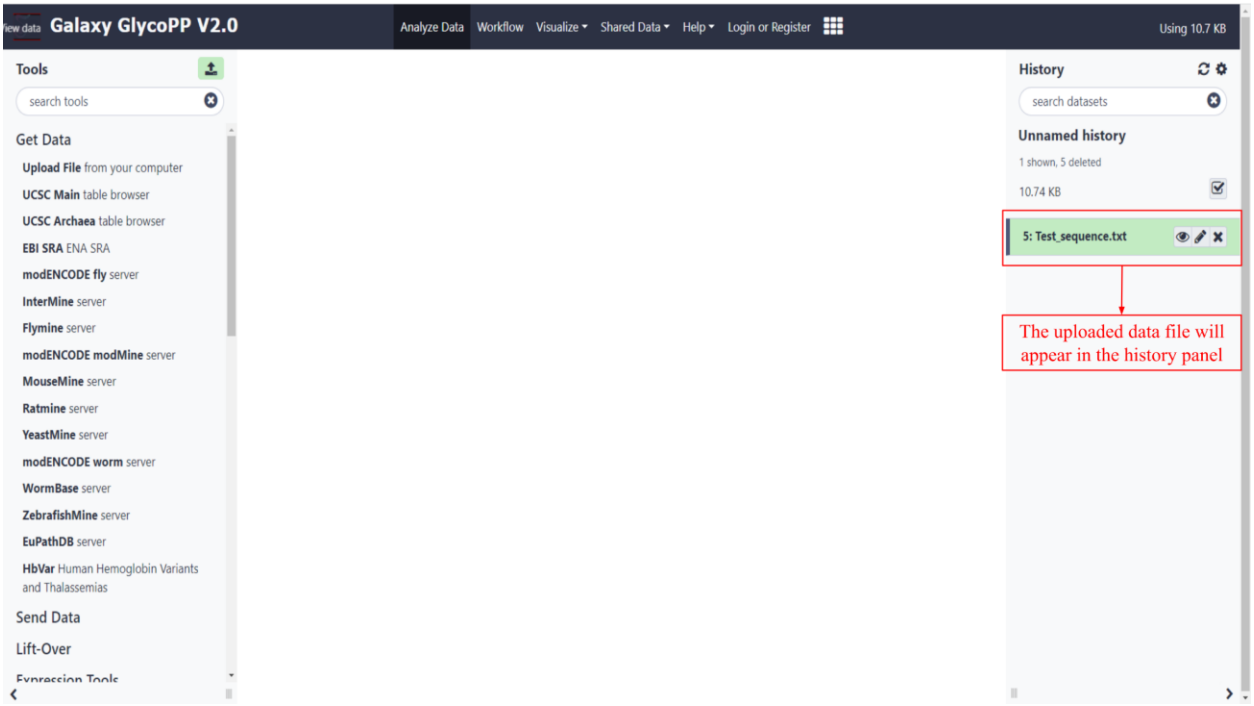


Figure: 15 The user can view, edit or delete the uploaded data file from the history panel.

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.

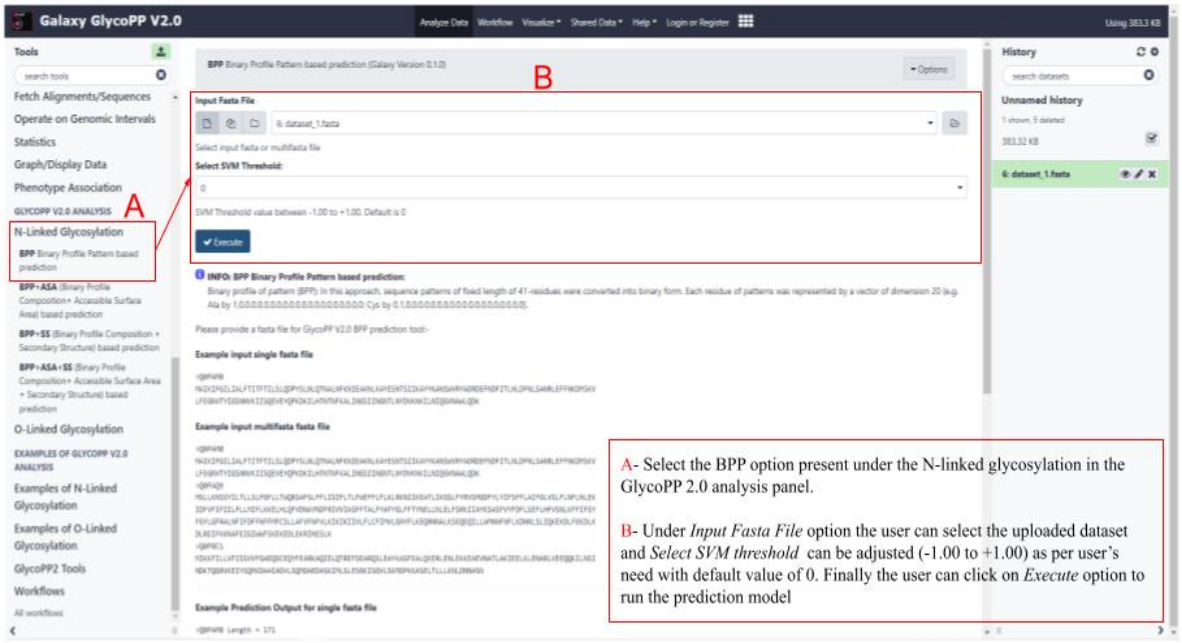


Figure 16: GlycoPP 2.0 model selection, Input file submission, SVM threshold selection and execution interface.

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.

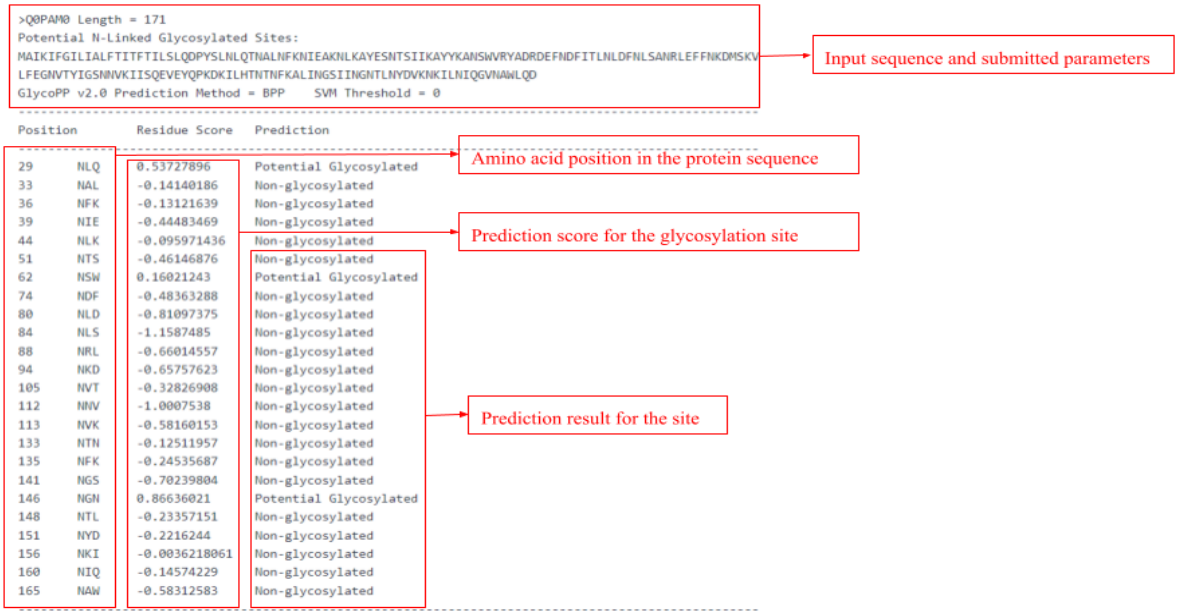


Figure 17: Screenshot of the GlycoPP 2.0 output page.

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