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 Nlinked 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 (<u>http://www.proglycprot.org/</u>).

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-<u>http://datascience.imtech.res.in/alkarao/glycopp2/index.html</u>. 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 linkhttp://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.

Galaxy GlycoPP	V2.0	Analy	ze Data Workflow Visualize+ Shared Data - Help - Login or Register 🇱	Using 5.9 KB
Tools	1			History C O
search tools	0	GlycoPP V2.0:	A webserver for glycosite prediction in prokaryotes	search datasets
Get Data				Unnamed history
Send Data		Home + Glycol	P V2.0 Shared Data • GlycoPP V2.0 Example Workflow • GlycoPP V2.0 User Manual •	6 deleted
Lift-Over				5.88 KB
Expression Tools	Overview	of GlycoPP V2.0		
Collection Operations	GlycoPP is covalanth	a webserver for predicting potential N-and O-glycosites in proi	aryotic protein sequence(s), where N-glycosite is an Asn residue and O-glycosite could be a serine or threonine residue having a glycan attached	This history is empty. You can load your own data or get data from an
Text Manipulation	GlycoPP	V2.0 is an updated version of our GlycoPP V1.0 (http://crdd.osd	d.net/raghava/glycopp) developed for highly accurate glycosylation prediction made available for the analysis of prokaryotic protein sequences on	external source
Convert Formats	the web b prokaryot	ased Galaxy Platform. GlycoPP prediction programmes are train- es as obtained from ProGlycProt V2.0 (http://www.proglycprot.	Id on the largest available and an extensive dataset of N-glycosites and O-glycosites extracted from experimentally characterized glycoproteins of org/l.	
Filter and Sort				
Join, Subtract and Group				
Fetch Alignments/Sequences	5 Dat	a collection & Processing	Features Extraction	
Operate on Genomic		[mail 1	41AA sequence pattern calculation Euli length protein calculations	
Intervals				
Statistics		PreGlycProt	CPP BPP PairkeD Patterns extracted of	
Graph/Display Data		Prese de la ferrar a la desana	SER CTD SARpred length 41 around N/S/T ASA	
Phenotype Association		Removed S-linked &	DPC (Q 0,1,2) PAAC (I 1,2,3) PSSM PPP	
GLYCOPP V2.0 ANALYSIS		protein		
N-Linked Glycosylation	(11)	W linked 148 O linked	4	
O-Linked Glycosylation	Bly	coprotein glycoprotein		
EXAMPLES OF GLYCOPP V2.0 ANALYSIS	1		SVM model generation	
Examples of N-Linked		Positive Dataset	CPP BPP SER CTD PPP DPC(Q 0,1,2) PAAC(I 0,1,2)	
Glycosylation	-		CPP+SS BPP+SS SER+SS CTD+SS PPP+SS DPC+SS PAAC+SS	
Examples of O-Linked	316	lycosite glycosite	CPP+ASA BPP+ASA SER+ASA CTD+ASA PPP + ASA DPC + ASA PAAC +ASS	
Glycosylation		41 AA Pattern	CPP+SS+ BPP+SS+ SER+SS+ CTD+SS+ ASA ASA ASA ASA ASA ASA ASA ASA ASA	

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.

Galaxy Glycol	P V2.0	Analyze Data Workflow Visualize * Shared Data * Help * Login or Register		Using 2.0 KE					
Tools	<u>±</u>	Conjoint Triad Descriptor (CTD)	* History	00					
search tools	0	The conjoint triad feature is sequence information for proteins. Thereiny amino add types are clustered into seven classes to construct the C-triad feature. First, protein sequences are encoded into a numerical search clustered into a numerical							
Set Data		tild type is performed to obtain a 343-dimensional numerical vector.							
Send Data		PSSM profile of patterns (PPP)	2 deleted						
ift-Over		The multiple sequence alignment information in the form of position specific scoring matrix (PSSM) has been used as input feature to develop this learning model. Each target sequence was scanned at Swiss-	1.96 KB						
xpression Tools		Prot to generate the alignment profiles or position specific scoring matrices (PSSM by PS-16LX) program. Three terrations of PS-16LX it were run for each protein with cut of te-value 0.001 thus generating the profile matrices. The PSSM contains probability of occurrence of each type of animo acid at each residue position of protein sequence. Finally we entrate PSSM contains probability of occurrence of each type of each type of the site of the site of the profile matrices. The PSSM contains probability of occurrence of each type of each type of the site of the site of the profile matrices.	0						
Collection Operations		type of amino acid of fixed length sequence patterns from full length sequence PSSM matrix that is calles PSSM profile of patterns.	This history is empty. Yo your own data or get of the second s	iu can <mark>load</mark> data from an					
ext Manipulation		Dipeptide Composition (DPC):	external source						
Convert Formats		As sequence patterns of fixed length of 41-residues, we consider gapped dipeptides composition, where DPC1_AA represents an amino acid, having a gap of order Q (Q=0.1 and 2), here we get the best netromenon at D=1							
ilter and Sort		personnaine as que s							
oin, Subtract and Group		Pseudo Amino Acid Composition (PAAC)							
etch Alignments/Sequer	ces	resulto amino acia composition using a discrete mode to tepreterra a protein yet writout completely loang ta sequence-order information. Here in this study, we estrat each in residue's impact in the subsequent residue's impact and a 3, got the best result at [= 3.							
Operate on Genomic Inte	vals	Secondary Structure (SS)							
itatistics		Previous studies on exkaryotic glycoproteins suggested that the probability of finding glycosite was higher at positions where there was a secondary structure change.							
Graph/Display Data		Accessible Surface Area (ASA)							
henotype Association		Surface accessibility is employed as another important feature because glycosylation has tenancy to occur at extracellular regions of proteins with the side chain of amino acid in the sequon exposed to the							
SLYCOPP V2.0 ANALYSIS		surface.							
N-Linked Glycosylation		Hybrid Approaches:							
D-Linked Glycosylation		We have obtained the ASA and SS from SARpred and PSIPRED prediction respectively which contains amino acid of fixed length sequence patterns from full length sequence on protein.							
EXAMPLES OF GLYCOPP V2.0 ANALYSIS		In view of the current understanding that glycosylation occurs on folded proteins in prokaryotes, we also provide hybrid models of above mention properties of protein sequence patterns in combination of ASA SS, and ASA+SS as shown in graphical abstract above.							
Examples of N-Linked Glycosylation									
Examples of O-Linked		Take an interactive tour of GlycoPP v2.0: Galaxy UI × History × Scratchbook ×							
GlycoPP2 Tools		Gycor+VLU is a protein gicosylation prediction webserver to accelerate the research. Gycor+VLU is developed by Gycor+VLU is developed by Gycor+VLU is a protein gicosylation prediction webserver to accelerate the research. Gycor+VLU is developed by Gycor+VLU is a protein gicosylation prediction webserver to accelerate the research. Gycor+VLU is developed by Gycor+VLU is a protein gicosylation prediction webserver to accelerate the research. Gycor+VLU is developed by Gycor+VLU is a protein gicosylation prediction webserver to accelerate the research. Gycor+VLU is developed by Gycor+VLU is a protein gicosylation prediction webserver to accelerate the research. Gycor+VLU is developed by Gycor+VLU is a protein gicosylation prediction webserver.							
	-	Developed & Hosted by: CSIR-Institute of Microbial Technology, Sector-39A, Chandigarh, India (160036). https://www.imtech.res.in							
		"best view resolution at TaZUX1000 px screen							
		Take an interactive tour of GlycoPP v2.0: Galaxy UI » History » Scratchbook »							

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.

- 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 dimension 20 Ala by а vector of (e.g. by Cys by
- 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 41residues, 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) (I) 1, 2 and 3 and got the best result at I = 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.

 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.

Galaxy GlycoPP V2.	Analyze Data Workflow Visualize* Shared Data * Help * Login or Register	Using 2.0 KB
Tools 📩		History C 🗘
search tools	BPP_example binary Profile Pattern based prediction example (Galaxy Version 0.1.1)	search datasets
THE HIG SOL	Preloaded example for input fasta sequence	Unnamed history
Join, Subtract and Group	> 009840	2 deleted
Fetch Alignments/Sequences	MAIKIFGIUALFTITFTILSLQDPYSUNLQTNAUNFKNIEAKNLKAYESNTSIIKAYYKANSWVRYADRDEFNDFTLNLDFNLSANRLEFFNKDMSKV	1.96 KB
Operate on Genomic Intervals		
Statistics		This history is empty. You can load
Graph/Display Data	Input fasta sequence (Allowed special character in fasta header line '- = _, () /+ * ^, ;?1[]: Invalid character will replaced by 'X')	your own data or get data from an
Diserting Association	Select SVM Threshold:	external source
Phenotype Association	•	
GLYCOPP V2.0 ANALYSIS	SVM Threshold value between -1.00 to +1.00. Default is 0	
N-Linked Glycosylation		
O-Linked Glycosylation	✓ Execute	
EXAMPLES OF GLYCOPP V2.0 ANALYSIS Examples of N-Linked Glycosylation 2 BPP example Rinary Profile Pattern	INFOR BPP Binary Profile Pattern based prediction: Binary profile of pattern (BPP): In this approach, sequence patterns of fixed length of 41-residues were converted into binary form. Each residue of patterns was represented by a vector of dimension 20 (e.g. Als by 1.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0	
based prediction example	Example input single fasta file	
BPP+ASA_example (Binary Profile Composition+ Accessible Surface Area) based prediction example	XQBANB MAICHELLIALUTTITTI SLQPPSULUZMALIN KULANEUTISTUKAVANIKAPANEPINETULUZMA SARLEFTIKOPISA LIEGNTYTSSIRKYTSQPPEQHCULMITIR ALUDISTUKATUNDANCULQQARALQDK	
BPP+SS_example (Binary Profile Composition + Secondary Structure) based prediction example	Example input multifiasta fasta file	
BPP+ASA+SS_example (Binary Profile Composition+ Accessible Surface Area + Secondary Structure) based prediction example Examples of O-Linked	INALAUKE I IT ILLUGA I IT ILLUGA IN AUKANA AUKA ILLUGA INALIAN INALAUKA ILLUGA INALAWA INALAUKA INALAUKA INALAUKA INALAUKA INAL	
Glycopp2 Toole datascience.imtech.res.in/alkarao/glycopp_v2/tool	INTERVITUVYTESINGEREGEREFERIKURGETERINESEENEKS	

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 Nlinked 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.

Galaxy GlycoPP V2.0	Analyze Data Workflow Visualize* Shared Data * Help * Login or Register	Using 2.0 KB
Tools 🛓	BBB avamate Sizes Backle Battern learned acceletion scenario (Police Maria	History C O
search tools	• Options	search datasets
Operate on Genomic Intervals	Preloaded example for input fasta sequence	Unnamed history
Statistics		2 deleted
Graph/Display Data		1.96 KB
Phenotype Association		This history is emoty. You can load
GLYCOPP V2.0 ANALYSIS	Input fasta sequence (Allowed special character in fasta header line '- = () / + * ^ . ;?! []'. Invalid character will replaced by 'X')	your own data or get data from an
N-Linked Glycosylation	Select SVM Threshold:	external source
O-Linked Glycosylation	• <u>3.b</u>	
EXAMPLES OF GLYCOPP V2.0	SVM Threshold value between - 1.00 to + 1.00. Default is 0	
ANALYSIS	✓ Execute 3.c 3.d	
Glycosylation		
BPP_example Binary Profile Pattern based prediction example	IN INFO BYP Binary Yorkine Yattern base prediction: Binary profile yattern Byp Annual Statement and Statement a	
BPP+ASA_example (Binary Profile Composition+ Accessible Surface	Please provide a fasta file for GlycoPP V2.0 BPP prediction tool-	
Area) based prediction example	Example input single fasta file	
BPP+SS_example (Binary Profile Composition + Secondary Structure) based prediction example	-XQRAVIB MILLTALTITTTILSUQPYSUULUMINUHVAILEANILLANESITSIIKAYVAANBANVAANBERIOPTTUULUMILSANULEFPIKUMSKY LIEANNYTSEMINKILSUVYRYNYKALLIMITUHVALTUSEITUKATUUNVALTUUQPANAULQK	
BPP+ASA+SS_example (Binary Profile Composition+ Accessible	Example input multifasta fasta file	
Surface Area + Secondary Structure)	XQP/W0	
Examples of Q-Linked	INALITUELIUS I III ILSUUMI SUUUMUUMUEMILEUNISIINKI MANDENIIKI MANDENIIKI KUNDENIIKI KUNDENII LEEDINTIESSINKIISEEVEN PUOLIELINTIIKALIMESTINIITUNNIKUUTUU ALUNTIKSINKEETTIKUNSIIV	
Glycosylation	xqmraqn ISLLIDISSYLTLLSLFGYLLTUQISAFSLFFLISIFLTLINEFFLFLKLRIDIIXEATLIKGSLPYRV9HOPYLYIFSFFLAIFQLVSLFLIFLINLEK	
GlycoPP2 Tools	IDYVIPIILVLUTE/LKREUKQVOMANDFRIVNIASFFTALFVAFVOLFFTNELUR, EUFSRITLAKSASFVYDVLSEFUDFVSNLKFFIFSV FOYLGFRAUMFIPFFTSFISLAFVFIFVLKIKIIVILFLFIDMLOMFLKEQONNALKSEQEQILLMMNIFIEKONNUSLIQUEKOLFEKOLK	
Workflows	DLREIFKONAFEIDTIMFSLEKEDLEKETINESLK XQPPCL	
All workflows	VIXKYTLLIVYTSSVVFGARQDCEQVYEARKAQTELQTREFORMQSLEAVKASFEALQKERLBILEXKEERVANTLAKTEELXLEINALVEEQQKTLIIGI NOKTOGRWEETVSSVXGAALDAV, SSVDAEDASKTIVLSLESRKTSGV, SSVDRXKASELTLLLONLDNMSN	
<		-

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.

Galaxy GlycoPP \	V2.0	Analyze Data Workflow Visualize - Shared Data - Help - Login or Register	Using	g 2.0 KB
Tools	±		History	00
search tools	0	BPP_example binary Profile Pattern based prediction example (Galaxy Version 0.1.1)	search datasets	0
Onomia en Canania Internela		Preloaded example for input fasta sequence	Unnamed history	
Operate on Genomic Intervals Statistics Graph/Display Data	S	>Q0PAM0 MAIKIFGILIALFTITFTILSLQDPYSLNLQTNALNFKNIEAKNLKAYESNTSIIKAYYKANSWVRYADRDEFNDFITL NLDFNLSANRLEFFNKDMSKV LFEGNVTYIGSNNVKIISQEVEYQPKDKILHTNTNFKALINGSIINGNTLNYDVKNKILNIQGVNAWLQDK	2 deleted 1.96 KB	
Phenotype Association GLYCOPP V2.0 ANALYSIS N-Linked Glycosylation	l	Input fasta sequence (Allowed special character in fasta header line ' - = () / + * ^ , : ? ! []'. Invalid character will replaced by 'X') Select SVM Threshold:	This history is empty. You can your own data or get data fr external source	load om an
O-Linked Glycosylation		0		
EXAMPLES OF GLYCOPP V2.0 ANALYSIS		SVM Threshold value between -1.00 to +1.00, Default is 0		
Examples of N-Linked Glycosylation		✓ Execute 6.a		
BPP_example Binary Profile Pattern based prediction example		1 INFO: BPP Binary Profile Pattern based prediction: Binary profile of pattern (BPP): In this approach, sequence patterns of fixed length of 41-residues were		
BPP+ASA_example (Binary Profile Composition+ Accessible Surface Are	ea)	converted into binary form. Each residue of patterns 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; Cys by 0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0).		
based prediction example	-	Please provide a fasta file for GlycoPP V2.0 BPP prediction tool:-		
•		· · · · · · · · · · · · · · · · · · ·	*	

Figure: 8 Submission execution option of GlycoPP 2.0.

Galaxy GlycoPP V2	.0	Analyze Data Workflow Visualize - Shared Data - Help - Login or Register					Using 3.9 KB		
Tools 🛓	0	5						History	C 🕈
search tools		Executed BPP_example a	na successtull	y added 1 job to	the queue			search datasets	0
Operate on Genomic Intervals Statistics		It produces 2 outputs: • 3: BPP_example • 4: BPP: Glycosylation			Unnamed history 2 shown, 2 deleted 3 92 KB	3			
Graph/Display Data		You can check the status of	of queued jobs	s and view the re	sulting dat	ta by refreshing t	he History		
Phenotype Association		completed successfully or	'error' if prob	lems were encou	intered.	ning to finished	п	4: BPP: Glycosylation	• # ×
GLYCOPP V2.0 ANALYSIS								3: BPP_example	● # ×
N-Linked Glycosylation				6.b				6.0	
O-Linked Glycosylation									
EXAMPLES OF GLYCOPP V2.0 ANALYSIS									
Examples of N-Linked									
Glycosylation									
BPP_example Binary Profile Pattern based prediction example									
BPP+ASA_example (Binary Profile Composition+ Accessible Surface Area) based prediction example									
<								E	>

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.

Galaxy GlycoPP V2.0	Analyze Data	Workflow Visualize • Sh	ared Data 👻 Help 🔻	Login or Register			Using 3.9 KB
Tools 🛓	>Q0PAM0	Length = 171				History	00
search tools	>Q0PAM0	Length = 171				search datasets	0
· · · · ·	Potential N-Linked Glycosylated Sites:					Unnemed bistory	
Operate on Genomic Intervals	MAIKIFGILIALFITTFTTLSLQDPYSLNLQTNALNFK	NIEAKNLKAYESNTSIIKAYYKA	NSWVRYADRDEFNDFI	LNLDFNLSANRLEFFNKDMSKV		Onnamed history	
Statistics	GlycoPP v2.0 Prediction Method = BPP	KALINGSIINGNTLNYDVKNK SVM Threshold = 0	ILNIQGVNAWLQDK			2 shown, 2 deleted	R
Graph/Display Data						3.92 ND	-
Phenotype Association	Position	Residue	Score	Prediction		4: BPP: Glycosylation	● / ×
GLYCOPP V2.0 ANALYSIS	29	NLQ	-0.68692359	Non-glycosylated	<	3: BPP_example	• / ×
N-Linked Glycosylation	33	NAL	-0.72174881	Non-glycosylated	10		
In Elliked Olycosylution	36	NFK	-0.12207736	Non-glycosylated		7.a	J
O-Linked Glycosylation	39	NIE	-0.32470975	Non-glycosylated			
EXAMPLES OF GLYCOPP V2.0	44	NLK	-0.20398314	Non-glycosylated			
ANALYSIS	51	NTS	1.0002212	Potential Glycosylated			
Examples of N-Linked	62	NSW	-0.52280092	Non-glycosylated			
Glycosylation	74	NDF	-0.35368478	Non-glycosylated			
RDD exemple Risson Profile Pattern	80	NLD	-0.8578013	Non-glycosylated			
based prediction example	84	NLS	0.71732003	Potential Glycosylated			
DDD. ACA symmetry (Disease DesEls	88	NRL	-0.36008792	Non-glycosylated			
Composition+ Accessible Surface	94	NKD	-0,15063364	Non-glycosylated			
Area) based prediction example	105	NVT	0.63372566	Potential Glycosylated			
BPP+SS example (Binary Profile	112	NNV	-0.70692049	Non-glycosylated			
Composition + Secondary Structure)	113	NVK	-0.50830369	Non-glycosylated			
based prediction example	133	NTN	-0.17287851	Non-glycosylated			
BPP+ASA+SS_example (Binary Profile	135	NFK	-0.53119115	Non-glycosylated			
Composition+ Accessible Surface Area	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.



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

Galaxy GlycoPP	V2.0		Analyze Data	Workflow Visu	alize 🔹 Shared Data 🕶	Help 👻 Login or Register			Using 3.9 KE
Tools	t	>Q0PAM0-Position	>Q0PAM0-Residue	BPP-TO-Score	BPP-T0-Prediction			History	0 (
search tools	0	>Q0PAM0-Position	>Q0PAM0-Residue	BPP-TO-Score	BPP-T0-Prediction			search datasets	0
	-	51	NTS	1.0002212	Potential Glycosylated				
Filter and Sort		84	NLS	0.71732003	Potential Glycosylated			Unnamed history	
Join, Subtract and Group		105	NVI	0.63372566	Potential Glycosylated Potential Glycosylated			2 shown, 2 deleted	
Fetch Alignments/Sequences								3.92 KB	Ľ
Operate on Genomic Intervals	5							4: BPP: Glycosylation	⊛ # X
Statistics								4 lines, 1 comments	
Graph/Display Data								B B C	
Phenotype Association							/	1.>Q0PAM0-Position 2.>Q0PAM0	-Residue 3.BPP
GLYCOPP V2.0 ANALYSIS								>Q0PAH0-Position >Q0PAH0-P	lesidue BPP-Ti
N-Linked Glycosylation						Developed wine datails		51 NIS 84 NLS	0.717
O-Linked Glycosylation						and Run this job again	70	105 NVT	0.633
EXAMPLES OF GLYCOPP V2.0						options		(•
Examples of N-Linked								3: BPP_example	● / ×
Glycosylation	- 1							format: tsv. database: ?	
BPP_example Binary Profile Pattern based prediction example								802	
BPP+ASA_example (Binary Profile Composition+ Accessible Surface Area) based prediction example								1.2089AN0 >Q09AN0 Potential N-Linked Glycosylat	ed Sites:
BPP+SS_example (Binary Profile								LFEGNUTYIGSNIVKIISQEVEYQPKDKI	LHTNTNFKALINGS1
	1.1							1	

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.

Salaxy GlycoPP V2.0	Analyze Data Workflow Visualize * Shared Data * Helo * Lonin or Register	Us	ing 7.3 KB
Tools	Download from web or upload from disk 6.1.a	History	00
search tools	Regular Composite Collection Rule-based	search datasets	8
Get Data		Unnamed history	
Upload File from your computer	×	3 shown, 2 deleted	
UCSC Main table browser	Name Size Type Genome Settings Status	7.33 KB	2
UCSC Archaea table browser	Test_sequence.txt 3.4 KB Auto-det Q Additional S 🔅 100%	St. Test companyon but	
EBI SRA ENA SRA		5: rest_sequence.txt	Ø .
modENCODE fly server		4: BPP: Glycosylation	0 X
InterMine server	Uploaded Uploaded Upload	4 lines, 1 comments	
Flymine server	progress	format: tsv, database: ?	
modENCODE modMine server		B 0 2	
MouseMine server		1.>Q0PAM0-Position 2.>Q0PAM0-Resid	due 3.8PP
Ratmine server		>Q0PAH0-Position >Q0PAH0-Residue	1.000
YeastMine server		84 NLS	0.717:
modENCODE worm server	Type (set all): Auto-detect Q Genome (set all): Additional S y 6,1,c	105 NVT	0.633:
WormBase server		141 NG5	0.731
ZebrafishMine server	6.1.b Choose local file @ Paste/Fetch data Pause Reset Start Close	3: RPP example	
EuPathDB server		22 lines 1 comments	
HbVar Human Hemoglobin Variants and Thalassemias		format: tsv, database: ?	
Send Data		802	
Lift-Over		1.3Q8PAH0 3Q8PAH0	
Fynression Tools		Potential N-Linked Glycosylated Si	(es: 🗸

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

option.

w data Galaxy GlycoPP	V2.0	Analyze Data Workflow Visualize ▼ Shared Data ▼ Help ▼	Login or Register
Tools	<u>±</u>		History
search tools	8		search datasets
Get Data			Unnamed history
Upload File from your computer			1 shown, 5 deleted
UCSC Main table browser			10.74 KB
UCSC Archaea table browser			
EBI SRA ENA SRA			5: Test_sequence.txt
modENCODE fly server			
InterMine server			
Flymine server			The uploaded da
modENCODE modMine server			appear in the his
MouseMine server			
Ratmine server			
YeastMine server			
modENCODE worm server			
WormBase server			
ZebrafishMine server			
EuPathDB server			
HbVar Human Hemoglobin Variants and Thalassemias	5		
Send Data			
Lift-Over			
Expression Tools	-		

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

 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.

 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.

>Q0PA Poten MAIKI LFEGN Glyco	>Q0PAM0 Length = 171 Potential N-Linked Glycosylated Sites: MAIKIFGILIALFTITFTILSLQDPYSLNLQTNALNFKNIEAKNLKAYESNTSIIKAYYKANSWVRYADRDEFNDFITLNLDFNLSANRLEFFNKDMSKV LFGNVTISOSUWXKIISQEVVQPKDKILHTNTNFKALINGSIINGNTLNYDVKNKTLNIQGVNANLQD GlycoPP v2.0 Prediction Method = BPP SVM Threshold = 0							
Posit	ion	Residue Score	Prediction					
29	NLO	0.53727896	Potential Glycosylated	Amino acid position in the protein sequence				
33	NAL	-0.14140186	Non-glycosylated					
36	NEK	-0.13121639	Non-glycosylated					
39	NIE	-0.44483469	Non-glycosylated					
44	NLK	-0.095971436	Non-glycosylated	Prediction score for the glycosylation site				
51	NTS	-0.46146876	Non-glycosylated					
62	NSW	0.16021243	Potential Glycosylated					
74	NDF	-0.48363288	Non-glycosylated					
80	NLD	-0.81097375	Non-glycosylated					
84	NLS	-1.1587485	Non-glycosylated					
88	NRL	-0.66014557	Non-glycosylated					
94	NKD	-0.65757623	Non-glycosylated					
105	NVT	-0.32826908	Non-glycosylated					
112	NNV	-1.0007538	Non-glycosylated	Pradiction result for the site				
113	NVK	-0.58160153	Non-glycosylated	Prediction result for the site				
133	NTN	-0.12511957	Non-glycosylated					
135	NEK	-0.24535687	Non-glycosylated					
141	NGS	-0.70239804	Non-glycosylated					
146	NGN	0.86636021	Potential Glycosylated					
148	NTL	-0.23357151	Non-glycosylated					
151	NYD	-0.2216244	Non-glycosylated					
156	NKI	-0.0036218061	Non-glycosylated					
160	NIQ	-0.14574229	Non-glycosylated					
165	NAW	-0.58312583	Non-glycosylated					

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

