

Structural bioinformatics

GLYCO: a tool to quantify glycan shielding of glycosylated proteins

Myungjin Lee ^{1,*}, Mateo Reveiz¹, Reda Rawi ¹, Peter D. Kwong^{1,2} and Gwo-Yu Chuang^{1,*}

¹Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA and ²Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received on August 7, 2021; revised on October 28, 2021; editorial decision on November 14, 2021; accepted on November 16, 2021

Abstract

Motivation: Glycans play important roles in protein folding and cell–cell interactions—and, furthermore, glycosylation of protein antigens can dramatically impact immune responses. While there have been attempts to quantify the glycan shielding or coverage of a protein surface, none of the publicly available tools analyzes glycan shielding computationally at an atomistic level.

Results: Here, we developed an *in silico* approach, GLYCO (GLYcan COverage), to quantify the glycan shielding of a protein surface. The software provides insights into glycan-dense/sparse regions of the entire protein surface or a subset of the protein surface. GLYCO calculates glycan shielding from a single coordinate file or from multiple coordinate files, for instance, as obtained from molecular dynamics simulations or by nuclear magnetic resonance spectroscopy structure determination, enabling analysis of glycan dynamics. Overall, GLYCO provides fundamental insights into the glycan shielding of glycosylated proteins.

Availability and implementation: GLYCO is freely available at GitHub (<https://github.com/myungjinlee/GLYCO>).

Contact: gwo-yu.chuang@nih.gov

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Glycans have diverse biological roles such as intercellular signaling, glycoprotein folding, biological masks (Dumville and Fry, 2000; Roth *et al.*, 1979; Schauer, 1985; Varki, 2017). In particular, glycan shielding can be a major factor in understanding protein–protein interactions as glycans can block protein binding and recognition. Especially in vaccine research, glycan shielding can substantially alter immune responses as it may markedly impede antibody–antigen recognition, whereas glycan holes generated by glycan removal can increase immunogenicity (Zhou *et al.*, 2017). Therefore, quantification of glycan shielding is crucial to glycoprotein related research.

Several studies have attempted to quantify and visualize the glycan shielding of protein surfaces (Berndsen *et al.*, 2020; Lemmin *et al.*, 2017; Stewart-Jones *et al.*, 2016; Wagh *et al.*, 2018). In addition, a few tools and servers for glycan modeling and delineation of potential glycosylation sites have been developed such as GlyProt (Bohne-Lang and von der Lieth, 2005), Glycan Reader & Modeler in CHARMM-GUI (Park *et al.*, 2019), and NetNGlyc 1.0 Server (Gupta and Brunak, 2002). However, they do not provide insight into the impact of glycosylation on the protein surface and

quantification of glycan shielding through a computational approach at an atomistic level has not been achieved.

Here, we introduce GLYCO (GLYcan COverage), a software to quantify glycan shielding or coverage of the glycoprotein surface. We have applied GLYCO to calculate the correlation between antibody–antigen properties and glycan shielding (Lee *et al.*, 2021) and to evaluate glycan-depleted areas, which antibodies preferentially target (Cerutti *et al.*, 2021).

GLYCO counts the number of glycan atoms per entire surface protein residue or per user input residue. It works not only for a single Protein Data Bank (PDB; Berman *et al.*, 2000) file but also for multiple frames, e.g. from molecular dynamics (MD) simulations or nuclear magnetic resonance spectroscopy structure determination. Overall, GLYCO is a useful tool for identifying glycan shielding over protein surfaces and can be applied to many research areas where glycosylated proteins are involved.

2 Materials and methods

GLYCO quantifies glycan shielding of glycoprotein surfaces using four steps (Fig. 1A). Only heavy atoms (e.g. non-hydrogen atoms)

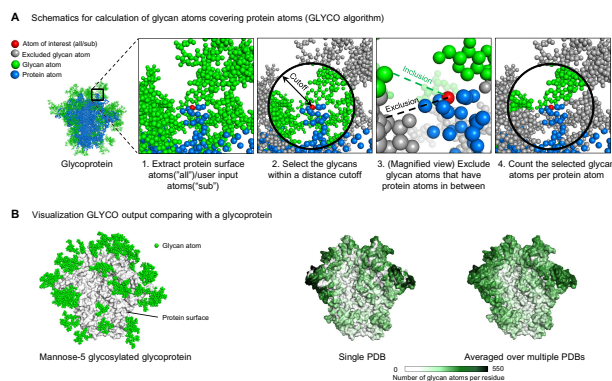


Fig. 1. Description of GLYCO, a glycan quantification protocol. (A) Schematics for calculation of glycan atoms covering protein surface atoms, the core algorithm for GLYCO. (B) Visualization of GLYCO output comparing with an actual glycoprotein. The example input is N-linked Mannose-5 BG505 N160K HIV-1 Env Trimer (left panel). The calculated glycan shielding of a single PDB is displayed in the middle panel and that of the multiple PDBs is in the right panel. The multiple frames are obtained from MD simulations, and the results were averaged over 300 frames (see [Supplementary Data](#) for more details)

of protein and glycan atoms are used in the analyses. The composition of glycan is not of importance to the calculation; thus, all types of glycans can be quantified by GLYCO (see [Supplementary Data](#) for workflow of GLYCO).

2.1 Step 1: extract protein surface residues

GLYCO executes FreeSASA ([Mitternacht, 2016](#)) to extract protein surface residues for analysis. The cutoff to define surface has been set to $30 \text{ \AA}^2$ with $1.4 \text{ \AA}$ as a probe radius by default; however, these can be altered by users.

2.2 Step 2: select glycans within a distance cutoff

GLYCO measures the distance from the atoms of protein surface residues that were selected from Step 1 to all glycan atoms. The glycan atoms beyond the defined distance cutoff are excluded from the calculations. The cutoff could be defined as the length of a glycan (as [Fig. 1B](#) using the longest length of glycans of the protein) but users can define it freely.

2.3 Step 3: exclude glycan atoms that have protein atoms in between

GLYCO defines glycan shielding as the number of glycan atoms covering a protein surface atom. Therefore, GLYCO excludes glycan atoms that are blocked by intervening protein regions, that do not actually provide shielding effect to the atom of the protein surface. The program generates vector equations of lines (r) between the vector of protein surface atoms (a) and the vector of selected glycan atoms (b)

$$r = a + t(b - a), \quad (1)$$

where $t \in [0, 1]$. Each discrete point along the vector line generates a cubic box from $-1 \text{ \AA}$ to $+1 \text{ \AA}$ on x , y , z and checks if any protein atom is within the box. If any protein atoms are in the box, the associated glycan atom is excluded from the calculation.

2.4 Step 4: count selected glycans per protein atom

The final step is to count the selected glycan heavy atoms associated with each protein surface atom within the distance cutoff. Steps 1–4 are repeated to generate glycan shielding for all protein surface atoms while removing redundant counts of glycan atoms in each protein residue.

3 Visualization

GLYCO outputs a PDB file that has glycan shielding values saved in the B-factor column (see [Supplementary Data](#)). Various molecular visualization programs such as PyMOL ([DeLano, 2002](#)) can display the glycan shielding based on the B-factors enabling for users to identify dense and sparse regions of glycan shielding ([Fig. 1B](#), middle and right) compared to the actual glycosylated protein ([Fig. 1B](#), left).

4 Conclusions

In this study, we developed GLYCO to quantify the glycan shielding of glycoproteins. This software calculates glycan shielding for the entire protein surface or a set of user-defined residues. The results can be visualized using protein structure visualization software. Considering the impact of glycans on proteins, such as the effect of viral proteins or immunogens, we believe that GLYCO will be a useful tool to quantify glycan shielding.

Acknowledgements

The authors thank J. Stuckey for assistance with figures, and members of the Structural Biology Section and Structural Bioinformatics Core, Vaccine Research Center, for discussions and comments on the manuscript. This work utilized the computational resources of the NIH HPC Biowulf cluster.

Funding

This work was supported by the Intramural Research Program of the Vaccine Research Center, National Institute of Allergy and Infectious Diseases. M.L. was supported by the Intramural AIDS Research Fellowship Program from the Office of AIDS Research, the Office of Intramural Training & Education and the Office of Intramural Research, National Institutes of Health.

Conflict of Interest: none declared.

References

- Berman, H.M. et al. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Berndsen, Z.T. et al. (2020) Visualization of the HIV-1 Env glycan shield across scales. *Proc. Natl. Acad. Sci. USA*, **117**, 28014–28025.
- Bohne-Lang, A. and von der Lieth, C.W. (2005) GlyProt: in silico glycosylation of proteins. *Nucleic Acids Res.*, **33**, W214–W219.
- Cerutti, G. et al. (2021) Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. *Cell Host Microbe*, **29**, 819–833.e7.
- DeLano, W. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA.
- Dumville, J.C. and Fry, S.C. (2000) Uronic acid-containing oligosaccharins: their biosynthesis, degradation and signalling roles in non-diseased plant tissues. *Plant Physiol. Biochem.*, **38**, 125–140.
- Gupta, R. and Brunak, S. (2002) Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput.*, **310–322**.
- Lee, M. et al. (2021) Extended antibody-framework-to-antigen distance observed exclusively with broad HIV-1-neutralizing antibodies recognizing glycan-dense surfaces. *Nat. Commun.*, **12**, 6470, 1–10.
- Lemmin, T. et al. (2017) Microsecond dynamics and network analysis of the HIV-1 SOSIP Env trimer reveal collective behavior and conserved microdomains of the glycan shield. *Structure*, **25**, 1631–1639.
- Mitternacht, S. (2016) FreeSASA: an open source C library for solvent accessible surface area calculations. *FI000Res*, **5**, 189.
- Park, S.J. et al. (2019) CHARMM-GUI glycan modeler for modeling and simulation of carbohydrates and glycoconjugates. *Glycobiology*, **29**, 320–331.
- Roth, M.G. et al. (1979) Polarity of influenza and vesicular stomatitis-virus maturation in MDCK cells—lack of a requirement for glycosylation of viral glycoproteins. *Proc. Natl. Acad. Sci. USA*, **76**, 6430–6434.

- Schauer,R. (1985) Sialic acids and their role as biological masks. *Trends Biochem. Sci.*, **10**, 357–360.
- Stewart-Jones,G.B.E. *et al.* (2016) Trimeric HIV-1-Env structures define glycan shields from clades A, B, and G. *Cell*, **165**, 813–826.
- Varki,A. (2017) Biological roles of glycans. *Glycobiology*, **27**, 3–49.
- Wagh,K. *et al.* (2018) Completeness of HIV-1 envelope glycan shield at transmission determines neutralization breadth. *Cell Rep.*, **25**, 893–908.
- Zhou,T. *et al.* (2017) Quantification of the impact of the HIV-1-glycan shield on antibody elicitation. *Cell Rep.*, **19**, 719–732.

GLYCO: a tool to quantify glycan shielding of glycosylated proteins

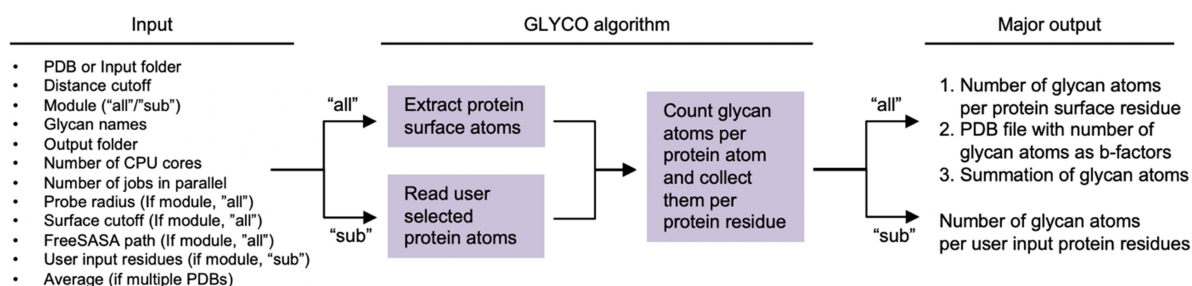
Myungjin Lee^{1,*}, Mateo Reveiz¹, Reda Rawi¹, Peter D. Kwong^{1,2} and Gwo-Yu Chuang^{1,*}

¹Vaccine Research Center, NIAID, National Institutes of Health, Bethesda, MD 20892, USA

²Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

*To whom correspondence should be addressed.

Supplementary data



Supplementary Fig. 1. Workflow to calculate glycan atoms for each protein surface residue for modules “all” and “sub” of a single PDB file or multiple PDBs.

1. Workflow of GLYCO

GLYCO requires several arguments depending on the number of input PDB (a single or multiple): distance cutoff, module type (“all” (**all** surface residues) or “sub” (**subset** residues)), glycan names in PDB, output folder name, number of CPU cores, number of jobs to run in parallel, cutoff for surface area (if module is “all”), probe radius cutoff (if module is “all”), path of the FreeSASA (Mitternacht, 2016) executable (if module is “all”), and list of input residues (if module is “sub”) (Supplementary Fig. 1).

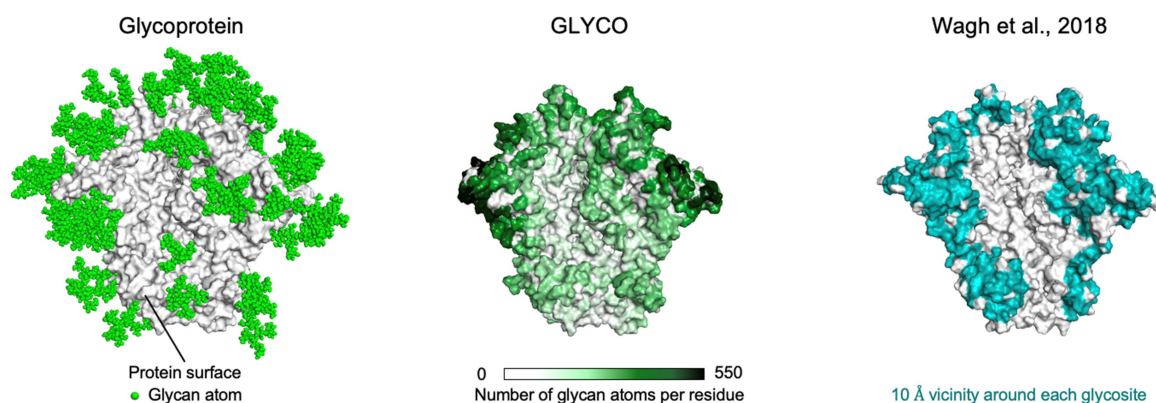
GLYCO is designed to read coordinates from the standard ATOM/HETATM record format of PDB (<http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html>). The distance cutoff can be defined by the users based on how far to include glycans from the protein surface. If users want to calculate number of glycan atoms per residue for all surface residues on a protein, module “all” should be entered. Users should specify glycan names in the PDB file such as NAG (*N*-acetyl- β -D-glucosamine), MAN (D-mannose), BGLN (*N*-acetyl- β -D-glucosamine). The name of output folder should be also entered as an argument. In addition, the number of CPU cores can be added to distribute a job into multiple cores. The surface area and probe radius cutoff can be entered, otherwise 30 Å² and 1.4 Å will be

used by default. Lastly, users should enter the path of the FreeSASA (Mitternacht, 2016) executable. Upon entering all arguments, GLYCO goes through the process that is described in Fig. 1A. The program then summates the number of glycan atoms of each protein atom in a residue while removing redundant counting of glycan atoms. Overall, the main output is a text file that includes residue name, chain ID, residue number and the number of glycan atoms. Additionally, a PDB file is generated containing the number of glycan atoms saved as B-factors. To compare the number of glycans per frame, the program also output the summation of the number of glycan atoms per input file which helps to track differences in glycans for each input file.

If users define module as “sub” to calculate number of glycan atoms of a subset of protein surface, the entering arguments should be the same as when module, “all”, but the users should enter a residue list as a file that has name of residue, chain ID, and residue number so that GLYCO counts glycan atoms only for those residues. Any redundant counting glycan atom is excluded within the residue set. For this module, the FreeSASA (Mitternacht, 2016) path, a surface cutoff and a probe radius do not need to be entered. The summation of the number of glycan atoms covering the residue set is the output and provided as a text file. More details to run GLYCO and example files are provided at <https://github.com/myungjinlee/GLYCO>

2. Features

2.1 Visualization and comparison with other method



Supplementary Fig. 2. Comparison between GLYCO and one of the other methods of glycan shielding analysis utilizing a single PDB. The left and middle panels are identical to what is shown in Fig. 1B left and middle. The right panel has the protein surface highlighted in cyan for the 10 Å vicinity around each glycosite, the method of glycan shielding analysis proposed by Wagh and colleagues (Wagh et al., 2018).

For module, “all”, users may visualize the degree of glycan shielding on the protein surface with “_bfactor.pdb” in the output folder, which has the glycan shielding value as B-factors of the PDB file.

The “bfactor.pdb” file can be displayed through molecular visualization programs such as PyMOL (DeLano, 2002) (Fig. 1B middle and right panel and Supplementary Fig. 2 middle panel). GLYCO result can be compared with one of other glycan shielding analysis shown in Supplementary Fig. S2 right panel, as the compared method is to define 10 Å vicinity around all glycan sequons as dense in glycan, otherwise sparse (Stewart-Jones et al., 2016).

2.2 Incorporation of multiple PDBs

Considering the conformationally flexible nature of glycans and proteins, it is more realistic to incorporate glycan dynamics when calculating glycan shielding.

Multiple frames, for instance generated by methods such as MD simulations with a glycan modeling server such as CHARMM-GUI (Park et al., 2019) or NMR spectroscopy, can be processed with the multi-frame module of GLYCO. Users can input multiple frames as an input folder. GLYCO will then perform glycan shielding analysis on to all provided structures. The jobs can be distributed in the users’ system by defining the number of processors per one job and number of submitting jobs. Also, by adding an averaging argument, the program can calculate averaged glycan shielding. Upon completion of the calculation, users may visualize the individual PDB or the averaged glycan shielding on to a certain PDB surface with averaged output (when module “all”, Fig. 1B right). The glycan dynamics through trajectory can be captured by averaging glycan atoms over multiple frames as shown in Fig. 1B right.

2.3 Implementation and computational cost

GLYCO is implemented in Python3.7.4. The software was developed and tested on Linux systems. Computation time for a single PDB file of a mannose-5 glycosylated HIV-1 N160K Env trimer with 29298 protein atoms (1884 protein residue) and 12480 glycan atoms (78 glycans) is approximately 13 mins (with the glycan cutoff of 23 Å, the surface cutoff of 30 Å², 1.4 Å as the probe radius, 48 processors with the system 3.0 GHz (Intel Xeon Gold 6136 CPU), module “all”). Using the module “sub”, it requires approximately less than 1 min considering 34 input residues on the same structure and identical parameters used for module “all”.

References

- Wagh, K. *et al.* (2018) Completeness of HIV-1 Envelope Glycan Shield at Transmission Determines Neutralization Breadth. *Cell Rep* **25** (4), 893-908.
- Mitternacht, S. (2016) FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res* **5**, 189.
- DeLano, W. The PyMOL Molecular Graphics System. DeLano Scientific; San Carlos, CA, USA: 2002. In.; 2002.

- Stewart-Jones, G.B.E. *et al.* (2016) 1 Trimeric HIV-1-Env Structures Define Glycan Shields from Clades A, B, and G. *Cell* **165** (4), 813-826.
- Park, S.J. *et al.* (2019) CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates. *Glycobiology* **29** (4), 320-331.