University of Dundee

Visualisation of experimentally determined and predicted protein N-glycosylation and predicted glycosylphosphatidylinositol anchor addition in Trypanosoma brucei.

Tinti, Michele; Ferguson, Michael A. J.

SOFTWARE TOOL ARTICLE

Visualisation of experimentally determined and predicted protein N-glycosylation and predicted glycosylphosphatidylinositol anchor addition in *Trypanosoma brucei*. [version 1; peer review: 3 approved]

Michele Tinti ⬥, Michael A. J. Ferguson ⬥

Wellcome Centre for Anti-Infectives Research (WCAIR), School of Life Sciences, University of Dundee, Dundee, Scotland, DD1 5HN, UK

Abstract

Background: *Trypanosoma brucei* is a protozoan parasite and the etiological agent of human and animal African trypanosomiasis. The organism cycles between its mammalian host and tsetse vector. The host-dwelling bloodstream form of the parasite is covered with a monolayer of variant surface glycoprotein (VSG) that enables it to escape both the innate and adaptive immune systems. Within this coat reside lower-abundance surface glycoproteins that function as receptors and/or nutrient transporters. The glycosylation of the *Trypanosoma brucei* surface proteome is essential to evade the immune response and is mediated by three oligosaccharyltransferase genes; two of which, TbSTT3A and TbSTT3B, are expressed in the bloodstream form of the parasite.

Methods: We processed a recent dataset of our laboratory to visualise putative glycosylation sites of the Trypanosoma brucei proteome. We provided a visualisation for the predictions of glycosylation carried by TbSTT3A and TbSTT3B, and we augmented the visualisation with predictions for Glycosylphosphatidylinositol anchoring sites, domains and topology of the Trypanosoma brucei proteome.

Conclusions: We created a web service to explore the glycosylation sites of the Trypanosoma brucei oligosaccharyltransferases substrates, using data described in a recent publication of our laboratory. We also made a machine learning algorithm available as a web service, described in our recent publication, to distinguish between TbSTT3A and TbSTT3B substrates.

Keywords
Trypanosoma brucei, proteomics, glycobiology, N-glycosylation, glycosylphosphatidylinositol, oligosaccharyltransferase, OST, prediction

Open Peer Review

Approval Status ✔ ✔ ✔

	1	2	3
version 1 31 Jan 2022	✔ view	✔ view	✔ view

1. **James D. Bangs**, University at Buffalo (SUNY), Buffalo, USA

2. **Michael P. Barrett** ⬥, University of Glasgow, Glasgow, UK

3. **Emma M. Briggs** ⬥, University of Edinburgh, Edinburgh, UK
University of Glasgow, Glasgow, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Michael A. J. Ferguson (m.a.j.ferguson@dundee.ac.uk)

Author roles: Tinti M: Data Curation, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Ferguson MAJ**: Conceptualization, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Introduction

The protozoan parasite *Trypanosoma brucei* is transmitted to humans by the tsetse fly (Glossina species), which is found only in sub-Saharan Africa[1]. The parasite replicates as procyclic form (PCF) in the tsetse fly midgut and some differentiate during migration to the salivary glands to replicating epimastigote forms. The latter differentiate into non-dividing metacyclic trypomastigotes that establish the mammalian host infection during a tsetse vector bloodmeal. Once in the host, the parasites differentiate into replicating, slender trypomastigotes and some of these differentiate into non-dividing stumpy forms that are adapted for survival and differentiation into procyclic forms once ingested by the vector[1]. Left untreated in the human host, the parasites invade the central nervous system causing neurological symptoms, coma and death[1]. The majority of experimental data on *T. brucei* have been obtained from either cultured versions of the bloodstream form (BSF), or BSF cells from rodent infections, and/or from the cultured procyclic form (PCF) of the parasite.

Like all eukaryotes, *T. brucei* modifies most proteins that enter its secretory pathway through glycosylation. Since cell surface glycoproteins are at the interface between the cell and its environment, they often play central roles in eukaryotic cell biology; *T. brucei* is no exception. The BSF relies on a surface coat made of glycosylphosphatidylinositol (GPI) anchored and *N*-glycosylated variant surface glycoprotein (VSG) to evade the host innate immune system and the acquired immune system through antigenic variation[2]. The BSF also expresses other lower abundance glycoproteins including but not restricted to: a novel VSG-like transferrin receptor (TfR)[2–4], a lysosomal/endosomal protein called p67[5], invariant surface (ISG) and endoplasmic reticulum (IGP) glycoproteins[6,7], a Golgi/lysosomal glycoprotein tGLP-1[8], a membrane-bound histidine acid phosphatase TbMBAP1[9], flagellar adhesion zone glycoproteins Fla1–3[10,11], a flagellar pocket/endosomal system haptoglobin-hemoglobin receptor (HpHbr)[12] and serum resistance antigen (SRA)[13], a complement factor H receptor (FHR)[14] and a metacyclic trypomastigote-specific ISG[15]. Some of these are metacyclic and/or BSF specific glycoproteins (eg. VSG, TfR, ISG, TbMAP1, HpHbr, SRA, FHR) while others are also common to PCF trypanosomes. PCF parasites also express unique glycoproteins including but not limited to: the abundant GPI-anchored procyclins, some of which are *N*-glycosylated[16,17], and a high-molecular weight glycoconjugate[18,19].

The GPI anchor structures of some BSF VSGs[20–23] and the TfR[24] have been solved, as have those of PCF procyclins[16]. All contain the conserved GPI core but the BSF GPIs contain sn-1,2-dimyristoylglycerol lipid and sidechains of up to 1 βGal and up to 5 αGal residues whereas the PCF procyclin GPIs are inositol-acylated and contain sn-1-acylglycerol lipid and sidechains of branched, *N*-acetyllactosamine and lacto-*N*-biose repeats capped with α2–3 sialic acid[16,25,26]. Expression of a BSF VSG gene in PCF cells resulted in PCF-type GPI anchor inositol-acylation and sidechain structure[27]. We therefore conclude that *T. brucei*

GPI anchors can be categorized as BSF- or PCF-type according th the lifecycle stage they are expressed in.

Several of the *N*-glycan structures expressed by BSF *T. brucei* have been solved and these include conventional oligomannose and biantennary complex structures as well as paucimannose and extremely unusual 'giant' poly-*N*-acetyl-lactosamine (poly-LacNAc) containing complex structures[28–32]. In contrast, only oligomannose *N*-glycans have been structurally described in wild type PCF trypanosomes[16,33]. Eukaryotic oligosaccharyltransferase (OST) enzymes responsible for N-glycosylation operate on asparagine residues in N-glycosylation sequon motifs of asparagine, any amino acid except proline, serine or threonine (N.^P[S/T]). We showed that two OST enzymes in BSF *T. brucei*, named TbSTT3A and TbSTT3B, have different acceptor and donor substrate specificities[33]. Thus, TbSTT3A first transfers Man5GlcNAc2 from Man5GlcNAc2-PP-dolichol to any sequons in acidic peptide environments and TbSTT3B transfers Man9GlcNAc2 from Man9GlcNAc2-PP-dolichol to all remaining sequons. The sites modified by TbSTT3A with bi-antennary Man5GlcNAc2 can be further processed to paucimannose structures and a wide array of complex N-glycan structures, while the sites modified by TbSTT3B with tri-antennary Man9GlcNAc2 can be maximally processed to tri-antennary Man5GlcNAc2; i.e., these sites are exclusively occupied by oligomannose N-glycans. Using this information, we were able to create a predictor to distinguish between N-glycosylation sequons preferentially modified by TbSTT3A, leading to paucimannose and/or complex N-glycans, or TbSTT3B, leading it oligomannose N-glycans[33]. Experimental proteomics data used to train the predictor exploited the sensitivity and resistance, respectively, of oligomannose and paucimannose/complex N-glycans. Removal of oligomannsoe glycans by endoglycosidase H leaves behind a single N-acetylglucosamine residue and thus marks relevant tryptic peptides with a 203 D mass-tag. The endoglycosidase H resistant paucimannose/complex N-glycans were subsequently removed with peptide N-glycosidase F in the presence of $H_2^{18}O$, leaving behind [^{18}O]aspartate in place of asparagine and thus marking relevant tryptic peptides with a 3 D mass-tag.

To facilitate the visualisation and analysis of putative *T. brucei* glycoproteins based on their predicted amino acid sequences, we have combined the prediction of N-terminal signal peptides (that are generally required for protein entry into the secretory pathway), C-terminal GPI addition signal peptides, N-glycosylation sequon (classified as experimentally determined and/or predicted TbSTT3A or TbSTT3B substrates) transmembrane and other protein domains. We have created a free to use web service incorporating all these features that we believe will be useful to the trypanosome research community.

Methods

We used the mass spectrometry data described in 33 and deposited at the PRIDE database[34] with accession numbers: PXD007267 and PXD007268 to extract the BSF glycoprotein sequons preferentially modified by TbSTT3A (and therefore expressing

complex and/or paucimannose N-glycans) or TbSTT3B (and therefore expressing oligomannose N-glycans). We also computed the ratio of the complex/paucimannose modifications as TbSTT3A modified sites / (TbSTT3A modified sites + TbSTT3A modified sites). Similarly, we computed the ratio of the oligomannose modifications as TbSTT3B modified sites / (TbSTT3B modified sites + TbSTT3A modified sites). We also collected transmembrane topology and signal peptide predictions using the Phobius website https://phobius.sbc.su.se/index.html[35] and GPI anchor site predictions using the big-PI Predictor available at https://mendel.imp.ac.at/gpi/gpi_server.html[36]. The machine-learning algorithm to distinguish the sites preferentially modified by TbSTT3B or TbSTT3A in BSF *T. brucei* is the same described in [33]. We further collected protein domain predictions using the CDART server[37]. The protein ids, sequences and descriptions were retrieved from TriTrypDB version 28[38]. TriTrypDB stores also user-based comments regarding the gene of interest and gene ontology (GO) annotation terms that were also retrieved and incorporated in the web application.

Implementation
We implemented a web server using the tornado python package version 4.3 (https://www.tornadoweb.org/en/stable/). The user interface was developed in javascript using bootstrap version 3.3.7, jquery version 3.1.1 and datatables version 1.10.11. The feature visualisation panel uses the neXtProt feature viewer package version 0.1.44[39]. The website is hosted at http://134.36.66.166:8070/home.

Operation
We recommend hosting the application on a web server with 1MB of RAM and 50GB of disk space. The application runs using the Tornado HTTPServer (https://www.tornadoweb.org/en/stable/guide/running.html). The application code can be

cloned from the git repository or downloaded from Zenodo[40]. After creating and activating a conda environment with the packages listed in requirments.txt[40], move to the application folder and start the Tornado HTTPServer with "python glyc_web_server.py"

Use cases
The user is presented with a responsive web application with two main components: a protein feature browser (Figure 1 and Figure 2) and a type of glycans prediction (Figure 3).

Protein feature browser
The protein feature browser can be queried with a protein identification number (Figure 1.1). After clicking the search button, the protein description and comments tab are updated (Figure 1.2 and Figure 1.3). The comment tab reports on the presence of: 1) a signal peptide, 2) the presence of occupied N-glycosylation sequons, as determined by mass spectrometry, and 3) the presence of a predicted GPI anchoring site. The peptide list tab (Figure 1.4) reports all the N.^P[S/T] sequons identified in the protein. It reports the peptide sequence (peptide) extracted from +/- 6 amino acid surrounding the central asparagine. The central asparagine is colour coded blue if predicted to be be modified by TbSTT3B, and therefore carry olgomannose N-glycans, or red if predicted to be be modified by TbSTT3A in BSF cells, and therefore carry paucimannose or complex N-glycans, as reported in the Prediction column. The table further reports the N-glycan occupied site position in the protein sequence (Site), the number of peptides detected by mass spectrometry indicating they were originally occupied by endoglycosidse H-resistant complex/paucimannose glycans (MS_complex / paucimannose), the number of peptides detected by mass spectrometry indicating they were originally occupied by endoglycosidse H-sensitive oligomannose glycans (MS_oligomannose). The

Figure 1. Web Application Layout. Screen shot of the upper half of the web application user interface. 1) Input text to query the web server with a protein identification number. 2) Text area reporting the protein description. 3) Text area reporting the presence of three protein features: Signal peptide, Glycosylation sites and GPI anchor. 4) Tab reporting the N-glycan peptide sequences identified in the protein sequence. 5) Search field for the peptide sequences. 6) Download buttons for the table listing the peptide sequences.
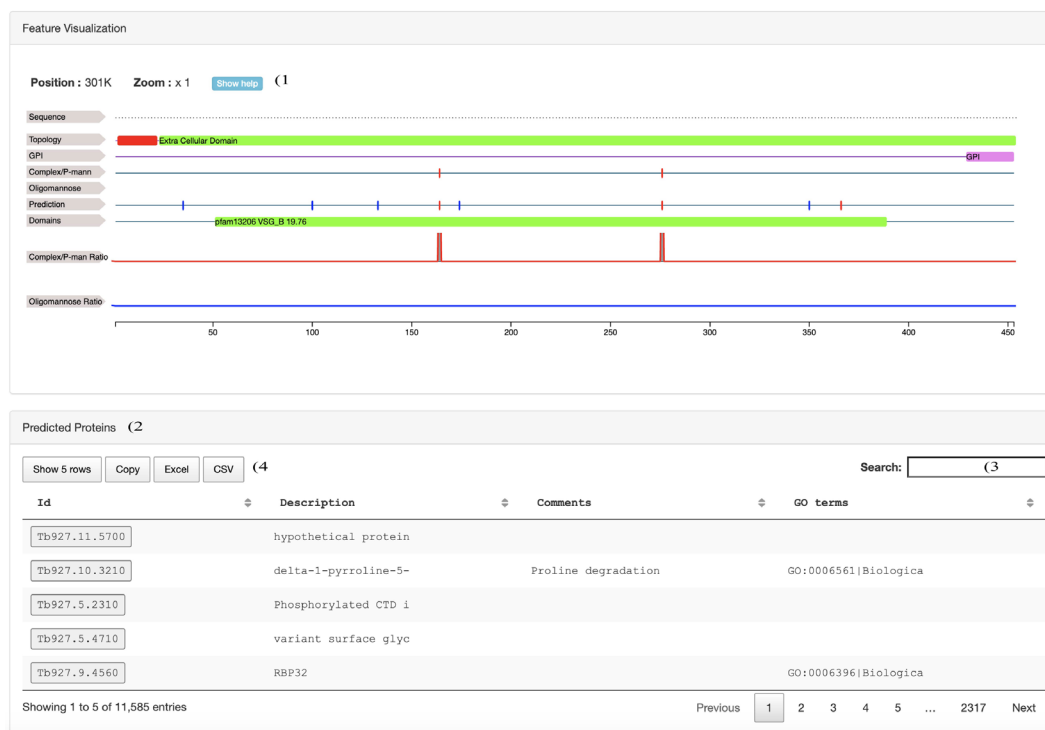
Figure 2. Web Application Layout. Screen shot of the bottom half of the web application user interface. 1) Feature visualization panel for the selected protein 2) Protein table listing the protein identification number available in the web application. 3) Search field for the protein table 4) Download buttons for the protein table.

table can be searched by peptide sequence or prediction type with the Search input field (Figure 1.5). The table can also be downloaded locally with the interaction buttons (Figure 1.6).

The protein identification number search button (Figure 1.2) also updates the visual protein sequence representation in the central part of the web page (Figure 2.1) reporting: 1) the protein sequence (Sequence), 2) the protein region predicted to be cleaved off after the addition of the GPI anchoring site (GPI), 3) The localisation of complex/paucimannose glycans identified by mass spectrometry, 4) The localisation of oligomannose glycans identified by mass spectrometry, 5) the CDART protein domain predictions, 6) the proportion of complex/paucimannose modifications and 7) the proportion of of the oligomannose modifications.

The full dataset hosted in the web application can be queried with the table at the bottom of the web application (Figure 2.2). The table can be searched using the search field (Figure 2.3) with the protein identification number (Id), gene description (Description), user-defined comments (Comments) and GO term annotations (GO term). The table can also be downloaded locally with the interaction buttons (Figure 2.4).

N-Glycan type prediction
The prediction link opens another user interface where it is possible to retrieve the prediction of a machine learning model trained to discriminate between sites preferentially modified by TbSTT3A (complex/paucimannose) or TbSTT3B (oligomannose) in BSF trypanosomes. The user can input a protein sequence in Fasta format (Figure 3.1), or an example sequence in Fasta format can be uploaded in the text input area by clicking on the Tb927.1.5100 protein id (Figure 3.2). After clicking on the Submit button (Figure 3.3) a results table is produced (Figure 3.4) reporting 1) the protein identification number (Prot), 2) The putative N.^P[S/T] sites in the protein as a peptide sequence (Seq) centred at the modified asparagine +/- 10 amino acids, 3) the predictor score (Score) and 4) the type of prediction (Prediction); Oligomannose glycans for TbSTT3B modified asparagine or Complex/Paucimannose glycans for TbSTT3A modified sites. The predictor was developed as a binary classifier for TbSTT3A modified sites using TbSTT3B modified as a negative set[33]. For this reason, a score close to 1 is indicative of a site preferentially modified by TbSTT3A. A score close to 0 is indicative of a site preferentially modified by TbSTT3B. A cut-off of 0.5 is used to determine if TbSTT3A or TbSTT3B is predicted to preferentially modify the asparagine.

Conclusions
We developed a web application to explore the glycosylation modifications mediated by TbSTT3A and TbSTT3B in the BSF proteome of *T. brucei*. It is important to re-emphasise that in wild type PCF *T. brucei*, only oligomannose N-glycans have been described and that this is largely controlled by suppression of

Figure 3. Protein Prediction page. Screen shot of the user interface to submit a protein sequence for predictions. 1) Input text area to copy\paste a protein sequence in FASTA format. 2) Submit button to start the prediction. 3) Text area to be populated with the prediction output.

TbSTT3A expression in that lifecycle stage. Thus, every occupied N-glycosylation sequon in wild type PCF cells is predicted to be of the oligomannose type.

It is also worth noting that the predictions that we present classify every asparagine in embedded in a N.^P[S/T] motif, even if it is biologically unlikely. For example, the predicted asparagine might reside in a protein that lacks an N-terminal signal peptide, or reside in a transmembrane region, in a signal peptide region or in the region excised after GPI modification of a protein. For this reason, we augmented our predictions with several visualisations of protein sequence properties (signal peptide, topology and GPI) predicted from other web services[35–37]). This should allow the interested user to evaluate both the type of glycan modifications and its biological relevance for the predicted sites.

Software availability

Source code available from: https://github.com/mtinti/glycosylation-server.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.5878703[40].

License: MIT.

Zenodo: mtinti/glycosylation-server: v0.1.

This project contains the following data:

- asap

  Python code to extract features from peptide sequence

- data

  Files to store pre-computed protein features

- models

  The model used for the glycosylation prediction

- scripts

  python code to parse protein features

- static

  javascript codes for the web server

- templates

  HTML code for the web server

- glyc_web_server.py

  Python code to start the web server

- predict_seq.py

  Python helper functions for the prediction page of the web server

- protein.py

  Python helper functions for the web server

- requirements.txt

  List of python packages to run the web server

References

1. Cox FE: **History of sleeping sickness (African trypanosomiasis).** *Infect Dis Clin North Am.* 2004; **18**(2): 231–45.
   PubMed Abstract | **Publisher Full Text**

2. Salmon D, Geuskens M, Hanocq F, *et al.*: **A novel heterodimeric transferrin receptor encoded by a pair of VSG expression site-associated genes in T. brucei.** *Cell.* 1994; **78**(1): 75–86.
   PubMed Abstract | **Publisher Full Text**

3. Trevor CE, Gonzalez-Munoz AL, Macleod OJS, *et al.*: **Structure of the trypanosome transferrin receptor reveals mechanisms of ligand recognition and immune evasion.** *Nat Microbiol.* 2019; **4**(12): 2074–2081.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. Mehlert A, Wormald MR, Ferguson MAJ, *et al.*: **Modeling of the N-glycosylated transferrin receptor suggests how transferrin binding can occur within the surface coat of *Trypanosoma brucei*.** *PLoS Pathog.* 2012; **8**(4): e1002618.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Peck RF, Shiflett AM, Schwartz FJ, *et al.*: **The LAMP-like protein p67 plays an essential role in the lysosome of African trypanosomes.** *Mol Microbiol.* 2008; **68**(4): 933–46.
   PubMed Abstract | **Publisher Full Text**

6. Jackson AP, Allison HC, Barry JD, *et al.*: **A cell-surface phylome for African trypanosomes.** *PLoS Negl Trop Dis.* 2013; **7**(3): e2121.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. Allison H, O'Reilly AJ, Sternberg J, *et al.*: **An extensive endoplasmic reticulum-localised glycoprotein family in trypanosomatids.** *Microb Cell.* 2014; **1**(10): 325–345.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

8. Lingnau A, Zufferey R, Lingnau M, *et al.*: **Characterization of tGLP-1, a Golgi and lysosome-associated, transmembrane glycoprotein of African trypanosomes.** *J Cell Sci.* 1999; **112 Pt 18**: 3061–70.
   PubMed Abstract

9. Engstler, M, Weise F, Bopp K, *et al.*: **The membrane-bound histidine acid phosphatase *Tb*MBAP1 is essential for endocytosis and membrane recycling in *Trypanosoma brucei*.** *J Cell Sci* 2005; **118**(Pt 10): 2105–18.
   PubMed Abstract | **Publisher Full Text**

10. LaCount DJ, Barrett B, Donelson JE: ***Trypanosoma brucei FLA1* is required for flagellum attachment and cytokinesis.** *J Biol Chem.* 2002; **277**(20): 17580–8.
    PubMed Abstract | **Publisher Full Text**

11. Woods K, Nic a'Bhaird N, Dooley C, *et al.*: **Identification and characterization of a stage specific membrane protein involved in flagellar attachment in *Trypanosoma brucei*.** *PLoS One.* 2013; **8**(1): e52846.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Higgins MK, Tkachenko O, Brown A, *et al.*: **Structure of the trypanosome haptoglobin-hemoglobin receptor and implications for nutrient uptake and innate immunity.** *Proc Natl Acad Sci U S A.* 2013; **110**(5): 1905–10.
    PubMed Abstract | **Publisher Full Text**

13. Zoll S, Lane-Serff H, Mehmood S, *et al.*: **The structure of serum resistance-associated protein and its implications for human African trypanosomiasis.** *Nat Microbiol.* 2018; **3**(3): 295–301.
    PubMed Abstract | **Publisher Full Text**

14. Macleod OJS, Bart JM, MacGregor P, *et al.*: **A receptor for the complement regulator factor H increases transmission of trypanosomes to tsetse flies.** *Nat Commun.* 2020; **11**(1): 1326.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

15. Casas-Sánchez A, Perally S, Ramaswamy R, *et al.*: **The crystal structure and localization of Trypanosoma brucei invariant surface glycoproteins suggest a more permissive VSG coat in the tsetse-transmitted metacyclic stage.** *bioRxiv.* 2018; 477737.
    Publisher Full Text

16. Treumann A, Zitzmann N, Hülsmeier A, *et al.*: **Structural characterisation of two forms of procyclic acidic repetitive protein expressed by procyclic forms of *Trypanosoma brucei*.** *J Mol Biol.* 1997; **269**(4): 529–47.
    PubMed Abstract | **Publisher Full Text**

17. Acosta-Serrano A, Cole RN, Mehlert A, *et al.*: **The procyclin repertoire of *Trypanosoma brucei*. Identification and structural characterization of the Glu-Pro-rich polypeptides.** *J Biol Chem.* 1999; **274**(42): 29763–71.
    PubMed Abstract | **Publisher Full Text**

18. Güther MLS, Lee S, Tetley L, *et al.*: **GPI-anchored proteins and free GPI glycolipids of procyclic form *Trypanosoma brucei* are nonessential for growth, are required for colonization of the tsetse fly, and are not the only components of the surface coat.** *Mol Biol Cell.* 2006; **17**(12): 5265–74.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

19. Güther MLS, Beattie K, Lamont DJ, *et al.*: **Fate of glycosylphosphatidylinositol (GPI)-less procyclin and characterization of sialylated non-GPI-anchored surface coat molecules of procyclic-form *Trypanosoma brucei*.** *Eukaryot Cell.* 2009; **8**(9): 1407–17.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

20. Ferguson MA, Homans SW, Dwek RA, *et al.*: **Glycosyl-phosphatidylinositol moiety that anchors *Trypanosoma brucei* variant surface glycoprotein to the membrane.** *Science.* 1988; **239**(4841 Pt 1): 753–9.
    PubMed Abstract | **Publisher Full Text**

21. Guther ML, Ferguson MA: **The microanalysis of glycosyl-phosphatidylinositol glycans.** *Methods Mol Biol.* 1993; **14**: 99–117.
    PubMed Abstract | **Publisher Full Text**

22. Mehlert A, Richardson JM, Ferguson MA: **Structure of the glycosylphosphatidylinositol membrane anchor glycan of a class-2 variant surface glycoprotein from** *Trypanosoma brucei*. *J Mol Biol.* 1998; **277**(2): 379–92.
PubMed Abstract | **Publisher Full Text**

23. Mehlert A, Sullivan L, Ferguson MA: **Glycotyping of** *Trypanosoma brucei* **variant surface glycoprotein MITat1.8.** *Mol Biochem Parasitol.* 2010; **174**(1): 74–7.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

24. Mehlert A, Ferguson MA: **Structure of the glycosylphosphatidylinositol anchor of the** *Trypanosoma brucei* **transferrin receptor.** *Mol Biochem Parasitol.* 2007; **151**(2): 220–3.
PubMed Abstract | **Publisher Full Text**

25. Pontes de Carvalho LC, Tomlinson S, Vandekerckhove F, *et al.*: **Characterization of a novel trans-sialidase of** *Trypanosoma brucei* **procyclic trypomastigotes and identification of procyclin as the main sialic acid acceptor.** *J Exp Med.* 1993; **177**(2): 465–74.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

26. Engstler M, Reuter G, Schauer R: **The developmentally regulated trans-sialidase from** *Trypanosoma brucei* **sialylates the procyclic acidic repetitive protein.** *Mol Biochem Parasitol.* 1993; **61**(1): 1–13.
PubMed Abstract | **Publisher Full Text**

27. Paturiaux-Hanocq F, Zitzmann N, Hanocq-Quertier J, *et al.*: **Expression of a variant surface glycoprotein of Trypanosoma gambiense in procyclic forms of** *Trypanosoma brucei* **shows that the cell type dictates the nature of the glycosylphosphatidylinositol membrane anchor attached to the glycoprotein.** *Biochem J.* 1997; **324**(Pt 3): 885–95.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. Zamze SE, Wooten EW, Ashford DA, *et al.*: **Characterisation of the asparagine-linked oligosaccharides from** *Trypanosoma brucei* **type-I variant surface glycoproteins.** *Eur J Bioche.* 1990; **187**(3): 657–63.
PubMed Abstract | **Publisher Full Text**

29. Zamze SE, Ashford DA, Wooten EW, *et al.*: **Structural characterization of the asparagine-linked oligosaccharides from** *Trypanosoma brucei* **type II and type III variant surface glycoproteins.** *J Biol Chem.* 1991; **266**(30): 20244–61.
PubMed Abstract

30. Atrih A, Richardson JM, Prescott AR, *et al.*: *Trypanosoma brucei* **glycoproteins contain novel giant poly-N-acetyllactosamine carbohydrate chains.** *J Biol*

Chem. 2005; **280**(2): 865–71.
PubMed Abstract | **Publisher Full Text**

31. Izquierdo L, Atrih A, Rodrigues JA, *et al.*: *Trypanosoma brucei* **UDP-glucose: glycoprotein glucosyltransferase has unusual substrate specificity and protects the parasite from stress.** *Eukaryot Cell.* 2009; **8**(2): 230–40.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

32. Acosta-Serrano A, O'Rear J, Quellhorst G, *et al.*: **Defects in the N-linked oligosaccharide biosynthetic pathway in a** *Trypanosoma brucei* **glycosylation mutant.** *Eukaryot Cell.* 2004; **3**(2): 255–63.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

33. Jinnelov A, Ali L, Tinti M, *et al.*: **Single-subunit oligosaccharyltransferases of** *Trypanosoma brucei* **display different and predictable peptide acceptor specificities.** *J Biol Chem.* 2017; **292**(49): 20328–20341.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

34. Perez-Riverol Y, Csordas A, Bai J, *et al.*: **The PRIDE database and related tools and resources in 2019: improving support for quantification data.** *Nucleic Acids Res.* 2019; **47**(D1): D442–D450.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

35. Kall L, Krogh A, Sonnhammer EL: **Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server.** *Nucleic Acids Res.* 2007; **35**(Web Server issue): W429–32.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

36. Eisenhaber B, Bork P, Eisenhaber F: **Prediction of potential GPI-modification sites in proprotein sequences.** *J Mol Biol.* 1999; **292**(3): 741–58.
PubMed Abstract | **Publisher Full Text**

37. Geer LY, Domrachev M, Lipman DJ, *et al.*: **CDART: protein homology by domain architecture.** *Genome Res.* 2002; **12**(10): 1619–23.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

38. Aslett M, Aurrecoechea C, Berriman M, *et al.*: **TriTrypDB: a functional genomic resource for the Trypanosomatidae.** *Nucleic Acids Res.* 2010; **38**(Database issue): D457–62.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

39. Paladin L, Schaeffer M, Gaudet P, *et al.*: **The Feature-Viewer: a visualization tool for positional annotations on a sequence.** *Bioinformatics.* 2020; **36**(10): 3244–3245.
PubMed Abstract | **Publisher Full Text**

40. mtinti: **mtinti/glycosylation-server: v0.1.** *Zenodo.* 2022.
http://www.doi.org/10.5281/zenodo.5878703

Open Peer Review

Current Peer Review Status: ✓ ✓ ✓

- -

Version 1

Reviewer Report 09 March 2022

https://doi.org/10.21956/wellcomeopenres.19514.r48799

✓ **Emma M. Briggs** 🆔

[1] Institute for Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Edinburgh, UK
[2] Wellcome Centre for Integrative Parasitology, University of Glasgow, Glasgow, UK

The authors present a web interface to allow users to explore the results of several prediction algorithms, including their own to predict *N*-glycosylation sites and discriminate TbSTT3A and TbSTT3B targets.

The web interface is functional and clear to use and will be useful to allow others to explore protein features.

The authors describe the use of Phobius, big-PI, and CDART for other predicted protein features. The parameters used for these can be included in the methods, even if all defaults were used. Which Phobius predictor was used (normal, constrained and homology supported predictions are all options on the referenced website).

Discussion of an experimentally validated prediction would be useful to assess the accuracy of predicted N-glycosylation sites, and other features. Reference 33 does not appear to have experimentally validated N-glycosylation predictions either.

Nevertheless, the web tool presented here has successfully enabled the community to make use of these predictions.

Typos:
  - "TbSTT3A modified sites / (TbSTT3A modified sites + TbSTT3A modified sites)"should these all be 3A?

  - "N-glycans" - N is sometimes in italics and sometimes not.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular parasitology, single cell transcriptomics, bioinformatics, trypansomes.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 28 February 2022

https://doi.org/10.21956/wellcomeopenres.19514.r48597

✓ **Michael P. Barrett** (iD)
Glasgow Polyomics, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

The submitted manuscript provides a succinct pointer towards a website and software tool allowing prediction of potential N-glycosylation sites related to the oligosaccharide transferases TbSST3A and TbSST3B in BSF T. brucei. The site is easily accessed and very fast. The code is accessible and can be downloaded too.

Very trivially, please correct the spelling of oligomannsoe 5 lines up in paragraph 4 of the Intro.

The one area I think important to consider though, as with any open access software offering, is how to future proof it. One solution would be to integrate into TritrypDB and it would be worth

contacting the EuPathDB team to discuss doing so. This could also help with another possible risk around changing accession numbers. Here TritrypDB version 28 has been used to bring query sequences directly through the algorithm. Linking the predictions to existing proteomics datasets that have previously been derived to seek for experimentally demonstrated N-glycosylation is a very positive addition to the software. However, this can bring problems where accession numbers change (and indeed one has to be careful of strain use and also isoform differences). For example, given the previous work Mehlert *et al.* 2012, PloS Pathogen[1]) on TfR N-glycosylation, this would be a good exemplar with which to see how the software performs. In haste, I pulled out the first visible ESAG6 and ESAG7 entries in tritrypDB and plugged those accession numbers in. For ESAG6, 5 predicted sites came out, but no MS hits were detected. For EASG7, 2 predicted sites came out, and no MS hits. As there are multiple isoforms of ESAG6 and ESAG 7 and sequences from multiple strains are present in tritrypDB matching the published and extracted sequences adds a layer of complexity. This particular example could be a good case study to include in the manuscript here to help users know how best to navigate the system.

For the uninitiated, it would also be useful to have a comment in the Introduction about the kinds of proteins likely to be N-glycosylated by TbSTT3A and 3B. I checked numerous transporters and enzymes in which I have a particular interest and many have predicted sites, but none show up in the MS datasets, presumably because it is membrane-destined proteins that originate in the ER that are substrates (hence the inclusion of information on the likely presence of signal peptides is useful, as are the other parts of information included in the visualisations). However, if, for example, potential N-glycosylation via TbSTT3A and B did become a feature in tritrypDB many false positive predictions will arise on possible sites, hence some clear qualifying prose here about protein types most likely to be true substrates would be helpful.

References
1. Mehlert A, Wormald MR, Ferguson MA: Modeling of the N-glycosylated transferrin receptor suggests how transferrin binding can occur within the surface coat of Trypanosoma brucei.*PLoS Pathog*. 2012; **8** (4): e1002618 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biochemical Parasitology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 02 February 2022

https://doi.org/10.21956/wellcomeopenres.19514.r48420

✔ **James D. Bangs**
Department of Microbiology and Immunology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo (SUNY), Buffalo, NY, USA

This short manuscript describes a newly developed website for in silico analyses of N-glycosylation in *Trypanosoma brucei.* The background for this is that the bloodstream form of this parasite adds distinct glycan structures (oligo- vs. pauci-mannose) based on two distinct oligosaccharyltransferases with different specificity for acidic or neutral/basic sequons. The underlying database is populated with experimentally determined site-specific data, and where this is absent, predictions based on the above-mentioned OST specificities. This reviewer spent several hours test driving the site with the highly glycosylated lysosomal transmembrane protein p67. It was very fun. This website will be useful for experienced glycobiologists and novices alike. There are several minor comments:

1. I believe that the passage on page 4 (column 1, first paragraph, lines 4-5) that reads: "TbSTT3A modified cites / (TbSTT3**A** modified sites + TbSTT3A modified sites)" should read: "TbSTT3A modified cites / (TbSTT3**B** modified sites + TbSTT3A modified sites)".

2. Suggestion: Add sequon pI on the 'features' list? Not necessary but might correlate with sites that are mixed pauci/oligomannose.

3. I had a different layout on the p67 'features visualization' section than the one you show in Fig 2. There are several extra lines, which I found useful, that are not in the example shown. To compare query Tb927.5.1810. May want to replace the example in the figure.

4. Zoom: I don't use double click on my mouse and I found it awkward to zoom in/out. I clicked on the sequon link or the position on the linear bars to zoom in (that was nice) but had to click on the original input query button to reset and zoom out. None of this is critical once you figure it out but perhaps this could be streamlined.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular and Cellular Parasitology, Cell Biology, Glycobiology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Feb 2022
Michele Tinti, School of Life Sciences, University of Dundee, Dundee, UK

1) Thanks for spotting this inconsistency. We will fix it in the next version of the paper.

2) I like this suggestion. I will add a polarity score to the feature panel; I'm thinking of using a sliding window of 5 amino acids.

3) I used a slighter older screenshot than intended. I will replace the figure in the next version of the paper.

4) Unfortunately, I don't think I can do any better for this functionality. I can zoom out with a two-finger click on my Mac laptop if it helps.

Competing Interests: No competing interests were disclosed.