FastProtein—an automated software for *in silico* proteomic analysis

Renato Simões Moreira[1,2], Vilmar Benetti Filho[2], Guilherme Augusto Maia[2], Tatiany Aparecida Teixeira Soratto[2], Eric Kazuo Kawagoe[2], Bruna Caroline Russi[1], Luiz Cláudio Miletti[3] and Glauber Wagner[2]

[1] Instituto Federal de Santa Catarina, Gaspar, Santa Catarina, Brazil
[2] Departamento de Microbiologia, Parasitologia e Imunologia, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, Brazil
[3] Centro de Ciências Agroveterinárias, Universidade do Estado de Santa Catarina, Lages, Santa Catarina, Brazil

ABSTRACT

Although various tools provide proteomic information, each tool has limitations related to execution platforms, libraries, versions, and data output format. Integrating data generated from different software is a laborious process that can prolong analysis time. Here, we present FastProtein, a protein analysis pipeline that is user-friendly, easily installable, and outputs important information about subcellular location, transmembrane domains, signal peptide, molecular weight, isoelectric point, hydropathy, aromaticity, gene ontology, endoplasmic reticulum retention domains, and N-glycosylation domains. It also helps determine the presence of glycosylphosphatidylinositol and obtain functional information from InterProScan, PANTHER, Pfam, and alignment-based annotation searches. FastProtein provides the scientific community with an easy-to-use computational tool for proteomic data analysis. It is applicable to both small datasets and proteome-wide studies. It can be used through the command line interface mode or a web interface installed on a local server. FastProtein significantly enhances proteomics analysis workflows by producing multiple results in a single-step process, thereby streamlining and accelerating the overall analysis. The software is open-source and freely available. Installation and execution instructions, as well as the source code and test files generated for tool validation, are available at https://github.com/bioinformatics-ufsc/FastProtein.

Subjects Bioinformatics, Computational Biology, Genomics
Keywords User-friendly proteomics, Docker, Web-based software, Proteomics

Submitted 30 April 2024
Accepted 24 September 2024
Published 31 October 2024

Corresponding author
Glauber Wagner,
glauber.wagner@ufsc.br

Academic editor
Armando Sunny

Additional Information and Declarations can be found on page 9

DOI **10.7717/peerj.18309**

INTRODUCTION

The complexity of high-throughput sequencing data and the need for reproducible analysis are challenges that require integrated workflows (*Wratten, Wilm & Göke, 2021*). Some workflow managers for bioinformatics include community-driven projects and workflow management systems. Galaxy (*The Galaxy Community et al.,, 2024*) and nf-core (*Ewels et al., 2020*) are examples of community-driven projects, while Nextflow (*Di Tommaso et al., 2017*) and Snakemake (*Mölder et al., 2021*) are script-based workflow management systems. Workflow managers and automated software facilitate reproducible and scalable data analysis. However, Galaxy is the only platform with a user-friendly interface and

a point-and-click feature to create workflows (*Wratten, Wilm & Göke, 2021*), while the remaining tools require the command-line interface.

Downstream analysis results in qualitative and quantitative features of proteins, which typically involve using several bioinformatics software packages in tandem but in a non-integrated workflow (*Chen et al., 2020*; *Jiménez-Munguía et al., 2018*). Although Blast2GO (*Conesa et al., 2005*) is an alternative tool widely used for functional annotation, it is closed sourced and requires a license. Proteomic analysis generates a considerable amount of computational data that require bioinformatics analysis (*Vaudel et al., 2016*).

FastProtein is an automated, user-friendly, and publicly available software that integrates functional annotation, database similarity search, and protein feature prediction to enable global proteomic profiling. Furthermore, the *in silico* results obtained through FastProtein can be used to characterize proteins of interest in search of biological insights.

MATERIALS & METHODS

Workflow

FastProtein uses a protein FASTA file as input to generate protein profiles. The workflow analysis begins by parsing and standardizing the input for different software (Fig. 1). Parsing and input validation are executed by BioJava (*Lafita et al., 2019*). Sequences with undetermined amino acids (represented by the letter "X") are invalid and removed from the initial dataset before execution.

The first step of FastProtein involves biochemical feature prediction, which provides attributes such as protein length, molecular mass (kDa), hydropathy, isoelectric point (PI), and aromaticity (*Lobry & Gautier, 1994*). These processes are managed and executed through BioJava (*Lafita et al., 2019*). Subsequently, FastProtein identifies the N-glycosylation and endoplasmic reticulum retention domains using the PROSITE (*Sigrist et al., 2013*) database entries PS00001 and PS00014, respectively.

WoLF PSORT (*Horton et al., 2007*) is used to predict the subcellular locations of eukaryotic organisms. Transmembrane site, signal peptide, and glycosylphosphatidylinositol (GPI)-anchored predictions are performed using TMHMM-2 (*Möller, Croning & Apweiler, 2001*), SignalP 5.0, (*Almagro Armenteros et al., 2019*), and PredGPI (*Pierleoni, Martelli & Casadio, 2008*), respectively. Additionally, the transmembrane domain and signal peptide are predicted using Phobius (*Käll, Krogh & Sonnhammer, 2004*).

Functional annotations are optional and performed using InterProScan (*Jones et al., 2014*). The outputs are merged and parsed to obtain Pfam (*Mistry et al., 2021*) and PANTHER (*Thomas et al., 2022*) domains, InterPro (IPR) annotations (*Jones et al., 2014*), and gene ontology (GO) terms (*Ashburner et al., 2000*). The sets of GO terms associated with the protein are determined by analyzing all databases using InterProScan. These terms are organized into a file that can be imported into the WEGO 2.0 (*Ye et al., 2018*) platform for complementary analysis. This platform is used to group, visualize, compare, and generate GO plots.

GO terms provide quantitative reports on molecular functions, cellular components, and biological processes. This process generates a file containing the GO terms (one line per protein, followed by GO terms in table-separated files).

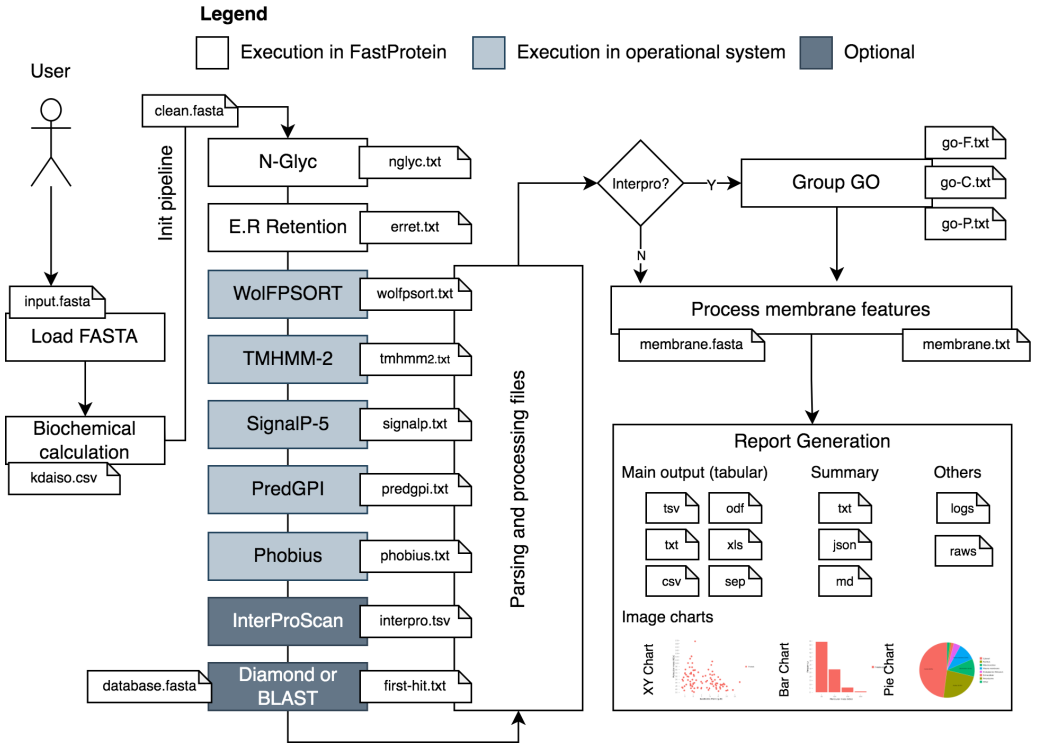

Figure 1 **FastProtein workflow.** A user-provided FASTA file are processed and submitted to third-party software that genarate outputs with GO terms, membrane evidence and other relevant graphs and files.
Full-size DOI: 10.7717/peerj.18309/fig-1

Another important step in this workflow is similarity analysis. FastProtein returns the best hit for each protein (with its identity and coverage percentage) using the BLASTp (*Camacho et al., 2009*) or DIAMOND (ultra-sensitive mode) (*Buchfink, Reuter & Drost, 2021*) algorithms. Local alignment using BLASTp is only available through the command-line interface (CLI). DIAMOND is the only local aligner available through the web server.

FastProtein provides six features to predict membrane protein: GPI-anchored, two predictions for transmembrane domains (TM, predicted *via* TMHMM-2.0; and PHOBIUS_TM, predicted *via* Phobius), subcellular localization predicted *via* WoLF PSORT (SL), GO, and IPR annotations. The presence of any one of these six features is sufficient to classify a protein as a membrane protein. Finally, a report file and a FASTA file are generated for proteins with membrane-related evidence.

Output files

FastProtein generates multiple outputs, including both quantitative and qualitative results, as well as FASTA files (Supplemental Information 1). Additionally, it produces an integrated histogram and scatter plot of molecular masses and isoelectric points, along with a bar chart depicting predicted subcellular localizations. The images are created at 300 DPI using Matplotlib (*Hunter, 2007*) and seaborn (*Waskom, 2021*).

Individual protein information is provided in tab-separated values (TSV), comma-separated values (CSV), plain text (TXT), XLS (Microsoft Excel), open document format

Table 1 Third-party software used by FastProtein.

Software/Version	Purpose	Reference
WoLF PSORT (0.1)	Subcellular location (for eukaryotes only)	*Horton et al. (2007)*
TMHMM (2.0c)	Transmembrane predictions of domain sites	*Möller, Croning & Apweiler (2001)*
SignalP (5.0b)	Signal peptide prediction and location of cleavage sites in protein	*Almagro Armenteros et al. (2019)*
InterProScan (5.61–93.0)	Functional predictions (Ontology terms)	*Jones et al. (2014)*
BioJava (7.0.0)	Bioinformatic support library	*Lafita et al. (2019)*
Phobius (1.01)	A combined transmembrane topology and signal peptide predictor	*Käll, Krogh & Sonnhammer (2004)*
PredGPI (202001)	GPI-Anchor Predictor	*Pierleoni, Martelli & Casadio (2008)*
BLASTp (2.10.0)	Sequence aligner for proteins	*Camacho et al. (2009)*
DIAMOND (2.0.7)	Sequence aligner for proteins	*Buchfink, Reuter & Drost (2021)*
Seaborn (0.13.2)	Python data visualization library	*Waskom (2021)*
Matplotlib (3.9.2)	Python library for creating static, animated, and interactive visualizations	*Hunter (2007)*

(ODF), and separated (SEP) file formats. SEP is a custom format similar to the ProtComp v9 (http://www.softberry.com/) output.

All generated files are stored in a temporary directory within the FastProtein installation directory, named using a universally unique identifier (UUID) created at the start of the run, which enables parallel execution. Upon completion, the temporary directory is renamed to the user-selected output directory (with 'fastprotein _results' as the default). In case of processing errors, the previously generated files can be reused by employing the <-cdt directory>command, which specifies the directory from which the files should be retrieved. This option is only available through the CLI mode. An execution log is saved in the output directory, and the logging level can be set in the CLI mode. Furthermore, the FastProtein Docker container functions as a comprehensive bioinformatics suite, featuring several pre-installed software packages (Table 1) that can be executed independently.

User friendly web-based interface

A web module was developed using Python (v3.9.2) and Flask (v2.2.3) to execute FastProtein within a Docker container. This interface enables users submit new FastProtein executions and monitor the progress of their tasks (Fig. 2) and visualize results through charts (Fig. 3A) and an interactive table (Fig. 3B). Additionally, it includes modules for managing databases, users, and permissions.

Computational infrastructure

A Debian-based Docker image (*Merkel, 2014*) is available at https://hub.docker.com/r/bioinfoufsc/fastprotein. This image is 900 MB (compressed) and includes an installation script for InterProScan, which is required for functional annotation (recommended). The third-party software and dependencies used are listed in Table 1 and the commands for each third-party software are detailed in Supplemental Information 2.

The installation guide, usage instructions, and the source code are available at https://github.com/bioinformatics-ufsc/FastProtein. FastProtein can be executed in two different ways: a web-based GUI and through the CLI from a local FastProtein Docker

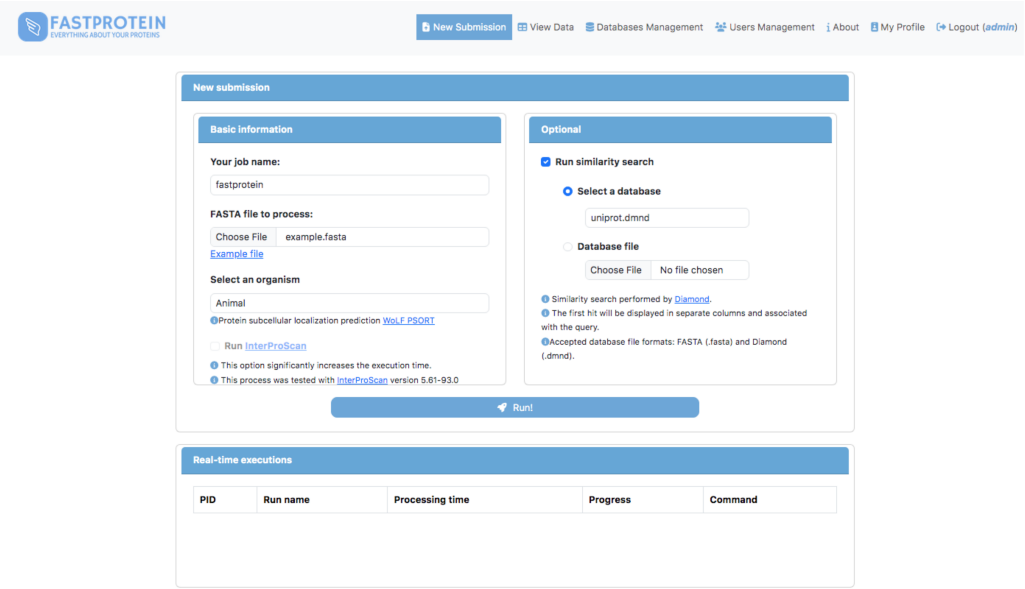

Figure 2 **Main GUI of the FastProtein web module deployed on a local server.**
Full-size DOI: 10.7717/peerj.18309/fig-2

container Users can also build a new FastProtein Docker image using the Dockerfile available at https://github.com/bioinformatics-ufsc/FastProtein.

Experiments

We used a subset of 125 proteins from the *Plasmodium malariae* proteome (UP000219813) to demonstrate the FastProtein workflow. This test run was performed through the CLI mode using DIAMOND (ultra-sensitive mode) and BLASTp for local alignment.

For proteome-wide analysis and performance benchmarking, we used the following proteomes (strain, Proteome ID) downloaded on April 2, 2023, from UniProt (*The UniProt Consortium et al., 2021*): *Plasmodium vivax* (strain Salvador I, UP000008333), *Trypanosoma brucei* (strain 927/4 GUTat10.1, UP000008524), *Cryptosporidium muris* (strain RN66, UP000001460), *Toxoplasma gondii* (strain ATCC 50861/VEG, UP000002226), *Aspergillus novofumigatus* (strain IBT 16806, UP000234474), and *Cyanidioschyzon merolae* (strain NIES-3377/10D, UP000007014). All proteomes were analyzed using DIAMOND and BLASTp (against the respective proteomes), and each proteome was run in triplicate. The WoLF PSORT dataset was set up to consider the closest organism, as the models are restricted to animals, plants, and fungi.

RESULTS

Only 124 proteins from the initial dataset were deemed eligible for analysis. Protein A0A1A8WBL6 had invalid sequences (represented by the letter "X") and was removed from the dataset. The total analysis time for this dataset, using DIAMOND, was 2 min and 58 s, with 5 s exclusively required by FastProtein, and the remaining time by third-party software. The total execution time was 7 min and 13 s, using BLASTp, with 3 min and

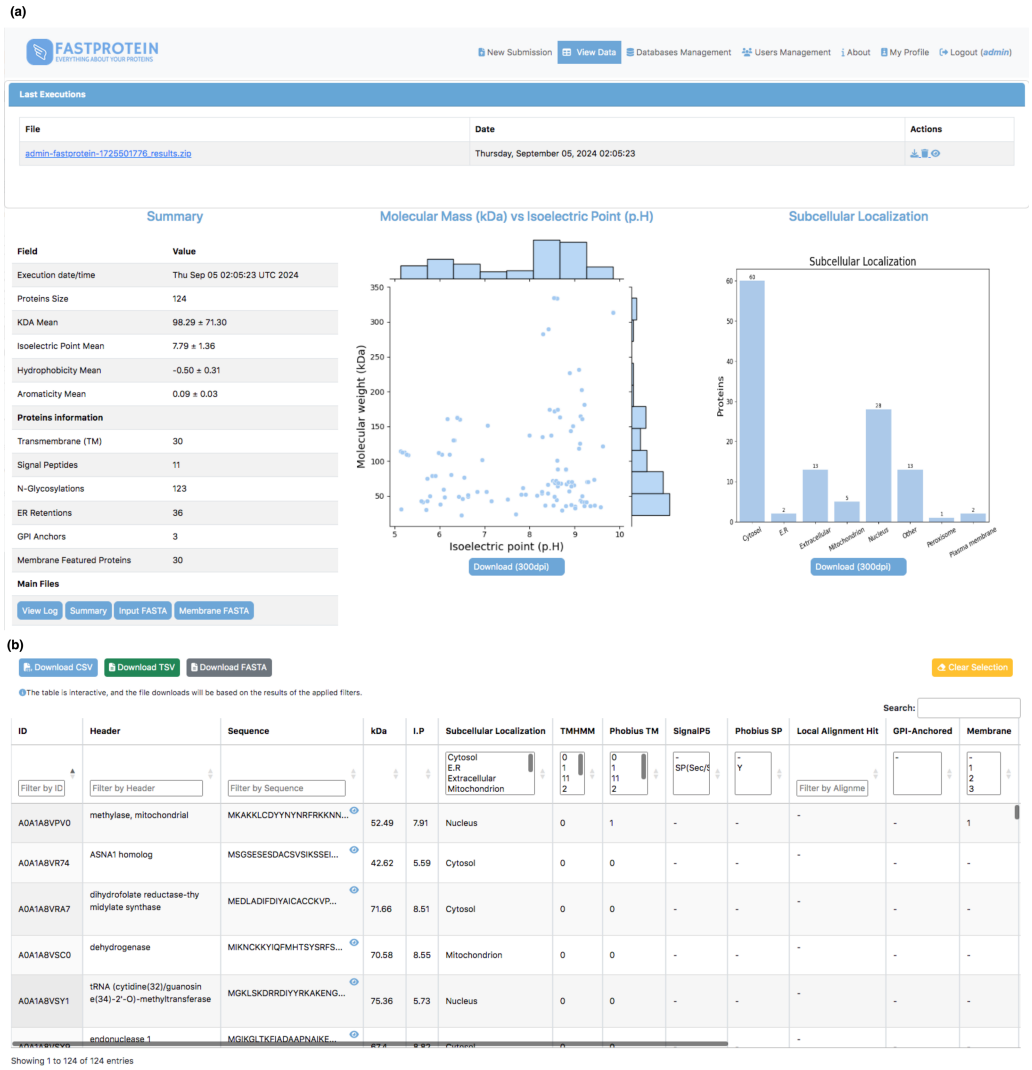

Figure 3 **Web interface for visualizing data generated by FastProtein.** (A) Executions are listed, and by clicking, the user is directed to the main view, which displays overall information about the run, including scatter plots of molecular weight *vs.* isoelectric point, as well as a bar chart for subcellular localization. (B) An interactive table where the user can filter data according to their needs, with the option to download the filtered data in CSV, TSV, or FASTA format.

Full-size DOI: 10.7717/peerj.18309/fig-3

49 s used for local alignment, 5 s by FastProtein, and the remaining by other third-party software.

The average molecular weight and isoelectric point in the *P. malariae* dataset were 98.29 ± 71.30 kDa and 7.79 ± 1.36, respectively. The distribution generated by FastProtein is shown in Fig. 3A. The average hydrophilicity and aromaticity were −0.50 ± 0.31 and 0.09 ± 0.03, respectively. Out of the analyzed proteins, 30 were predicted to contain transmembrane domains, three exhibited GPI anchoring, and an additional 30 were estimated to be membrane proteins. Furthermore, among the latter group, four exhibited multiple pieces of evidence supporting their localization to the membrane (GO, IPR,

PHOBIUS_TM, and TM). Among the 124 *P. malariae* proteins, 11 were predicted to have signal peptides and 36 had ER retention domains, with the KEEL and KNEL domains being the most frequently occurring, existing in seven proteins each. Of the 124 proteins, 123 had N-glycosylation domains, with the NNS domain occurring most frequently (116 proteins). The subcellular locations are shown in Fig. 3A, wherein cytosol and nucleus are the most frequent ones, with 60 and 28 proteins, respectively.

For the proteome-wide analysis, the fastest runtime was 28.85 ± 5.74 min for the *C. muris* dataset (3,930 proteins; 3,924 processed and six ignored) using DIAMOND. The slowest runtime was 262.35 ± 2.63 min for the *T. gondii* dataset (8,404 proteins) using BLASTp. The execution time decreased to 77.75 ± 8.90 min with DIAMOND as the local aligner. Even though *A. novofumigatus* had 3,076 more proteins than *T. gondii*, it took 97.42 ± 0.21 min to analyze it using BLASTp,and 67.71 ± 1.73 min using DIAMOND. The best result in terms of proteins analyzed per minute was obtained for the *A. novofumigatus* dataset with 170 proteins (using DIAMOND), whereas the worst was obtained for *T. gondii* with 32 proteins (using BLASTp). For all DIAMOND executions, the alignment was completed within a few seconds. For BLASTp, the fastest execution required 8.02 ± 0.05 min and the slowest required 193.06 ± 0.92 min. Considering our entire workflow, the local alignment was the only step with different execution times, which is considerably similar to previously reported results using different alignment algorithms (*Hernández-Salmerón & Moreno-Hagelsieb, 2020*).

The average execution times for each software used in the pipeline were as follows: WoLF PSORT (∼2 min), TMHMM-2.0 (∼8 min), SignalP5 (∼2 min), PredGPI (∼2 min), Phobius (∼11 min), InterProScan (∼25 min), DIAMOND (<1 min), and BLASTp (∼52 min). The average time required for the internal execution of FastProtein, file generation, conversions, and calculations were approximately 1 min. Both files generated from the subset of *P. malariae* proteins and proteome-wide rounds are available at https://github.com/bioinformatics-ufsc/FastProtein (including the intermediate files, which were removed during processing). All data analyses are presented in Supplemental Information 3.

DISCUSSION

FastProtein is a user-friendly and easy-to-install protein analysis pipeline tool that provides important information about protein datasets. FastProtein integrates calculations of molecular weight, isoelectric point, hydropathy, and aromaticity with predictions of subcellular location, transmembrane domains, signal peptide and GPI-anchor, GO, endoplasmic reticulum retention, and N-glycosylation domains. It also integrates results from InterProScan, PANTHER, Pfam, and alignment-based annotation searches. Additionally, the software provides a dataset of proteins with evidence of membrane localization, which is important for immunogenicity studies during vaccine development (*Cheng et al., 2021*; *Kis et al., 2018*) and diagnostic tests, such as ELISA (*De Haro-Cruz et al., 2019*; *Iha et al., 2022*) and western blotting (*Begum, Murugesan & Tangutur, 2022*; *Crescitelli, Lässer & Lötvall, 2021*; *Springhorn & Hoppe, 2019*).

FastProtein outputs files in formats that are widely used in the scientific community, including TSV, XLS, and ODF, as well as high-quality 300 DPI images, which is a widely used standard.

The total execution time of FastProtein depends on the InterProScan functional analysis and the local alignment method. By default, DIAMOND was selected due to its relatively faster execution time, although BLASTp is also available. The global median time for proteome-wide analyses was approximately 116 proteins per minute. This was increased to approximately 142 proteins per minute using DIAMOND, and approximately 90 proteins per minute. Thus, proteomic data from a large dataset can be quickly obtained using FastProtein. Considering differences in execution time and sensitivity, DIAMOND was chosen as the local aligner for the web server. DIAMOND can significantly decrease the alignment time (*Buchfink, Reuter & Drost, 2021*).

The only requirement for using the FastProtein software is the installation of Docker for local runs. FastProtein is an easy-to-use and viable tool for researchers with no background in bioinformatics because it provides a user-friendly interface similar to well-established software such as Blast2GO (*Conesa et al., 2005*), MEGA11 (*Tamura, Stecher & Kumar, 2021*), and MaxQuant (*Prianichnikov et al., 2020*). It also contributes to the initiatives that aim to democratize access to bioinformatics, such as the BioLib (https://biolib.com) and Galaxy (*The Galaxy Community et al., 2024*) projects.

CONCLUSIONS

FastProtein is a novel and user-friendly pipeline tool for proteomic data analysis that is available for small datasets and proteome-wide studies. Furthermore, it can be used through the CLI mode or a web interface. FastProtein accelerates proteomics analysis routines by generating multiple results in a one-step run. One of the limitations of FastProtein is that it does not yet integrate mass spectrometry data. However, the integration of both raw MS/MS data and data from other protein identification software through mass spectrometry is currently being implemented. The software is open-source and available at https://github.com/bioinformatics-ufsc/FastProtein, along with installation and execution instructions and test files generated for validation.

ACKNOWLEDGEMENTS

We are thankful to SeTIC (Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação) team from the Universidade Federal de Santa Catarina (UFSC) for the all computational infrastructure support and for hosting the FastProtein website.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by Santa Catarina Research Foundation (Fundação de Amparo à Pesquisa e Inovação of Santa Catarina, FAPESC, Santa Catarina, Brazil) and CAPES (Coordination for the Improvement of Higher Education Personnel, Brazil, Grant: 88881.311316/2018-01). Eric Kazuo Kawagoe, Guilherme Augusto Maia, and Vilmar Benetti Filho received scholarships from CAPES (Coordination for the Improvement of Higher Education Personnel, Brazil). Tatiany AT Soratto was a recipient of a scholarship from Santa Catarina Research Foundation (Fundação de Amparo à Pesquisa e Inovação of Santa Catarina, FAPESC, Santa Catarina, Brazil). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Santa Catarina Research Foundation (Fundação de Amparo à Pesquisa e Inovação of Santa Catarina, FAPESC, Santa Catarina, Brazil) CAPES (Coordination for the Improvement of Higher Education Personnel, Brazil, Grant: 88881.311316/2018-01).

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Renato Simões Moreira conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Vilmar Benetti Filho performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Guilherme Augusto Maia performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Tatiany Aparecida Teixeira Soratto performed the experiments, prepared figures and/or tables, and approved the final draft.
- Eric Kazuo Kawagoe performed the experiments, prepared figures and/or tables, and approved the final draft.
- Bruna Caroline Russi analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Luiz Cláudio Miletti conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Glauber Wagner conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The installation and execution instructions, the source code and test data generated for tool validation, are available at GitHub and Zenodo:

- https://github.com/bioinformatics-ufsc/FastProtein.

- Renato Simões, Vilmar Benetti Filho, & ekazuo. (2024). bioinformatics-ufsc/FastProtein: FastProtein-v1.0.0 (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.13838683.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.18309#supplemental-information.

REFERENCES

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, Von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology* **37**:420–423 DOI 10.1038/s41587-019-0036-z.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* **25**:25–29 DOI 10.1038/75556.

Begum H, Murugesan P, Tangutur AD. 2022. Western blotting: a powerful staple in scientific and biomedical research. *BioTechniques* **73**:58–69 DOI 10.2144/btn-2022-0003.

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**:366–368 DOI 10.1038/s41592-021-01101-x.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421 DOI 10.1186/1471-2105-10-421.

Chen X-L, Liu C, Tang B, Ren Z, Wang G-L, Liu W. 2020. Quantitative proteomics analysis reveals important roles of N-glycosylation on ER quality control system for development and pathogenesis in *Magnaporthe oryzae*. *PLOS pathogens* **16**:e1008355 DOI 10.1371/journal.ppat.1008355.

Cheng K, Zhao R, Li Y, Qi Y, Wang Y, Zhang Y, Qin H, Qin Y, Chen L, Li C, Liang J, Li Y, Xu J, Han X, Anderson GJ, Shi J, Ren L, Zhao X, Nie G. 2021. Bioengineered bacteria-derived outer membrane vesicles as a versatile antigen display platform for tumor vaccination *via* Plug-and-Display technology. *Nature Communications* **12**:2041 DOI 10.1038/s41467-021-22308-8.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**:3674–3676 DOI 10.1093/bioinformatics/bti610.

Crescitelli R, Lässer C, Lötvall J. 2021. Isolation and characterization of extracellular vesicle subpopulations from tissues. *Nature Protocols* **16**:1548–1580 DOI 10.1038/s41596-020-00466-1.

De Haro-Cruz MJ, Guadarrama-Macedo SI, López-Hurtado M, Escobedo-Guerra MR, Guerra-Infante FM. 2019. Obtaining an ELISA test based on a recombinant protein of *Chlamydia trachomatis*. *International Microbiology* **22**:471–478 DOI 10.1007/s10123-019-00074-4.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**:316–319 DOI 10.1038/nbt.3820.

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* **38**:276–278 DOI 10.1038/s41587-020-0439-x.

Hernández-Salmerón JE, Moreno-Hagelsieb G. 2020. Progress in quickly finding orthologs as reciprocal best hits: comparing BLAST, LAST, DIAMOND and MMseqs2. *BMC Genomics* **21**:741 DOI 10.1186/s12864-020-07132-6.

Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Research* **35**:W585–W587 DOI 10.1093/nar/gkm259.

Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering* **9**:90–95 DOI 10.1109/MCSE.2007.55.

Iha K, Tsurusawa N, Tsai H-Y, Lin M-W, Sonoda H, Watabe S, Yoshimura T, Ito E. 2022. Ultrasensitive ELISA detection of proteins in separated lumen and membrane fractions of cancer cell exosomes. *Analytical Biochemistry* **654**:114831 DOI 10.1016/j.ab.2022.114831.

Jiménez-Munguía I, Pulzova L, Kanova E, Tomeckova Z, Majerova P, Bhide K, Comor L, Sirochmanova I, Kovac A, Bhide M. 2018. Proteomic and bioinformatic pipeline to screen the ligands of *S. pneumoniae* interacting with human brain microvascular endothelial cells. *Scientific Reports* **8**:5231 DOI 10.1038/s41598-018-23485-1.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240 DOI 10.1093/bioinformatics/btu031.

Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology* **338**:1027–1036 DOI 10.1016/j.jmb.2004.03.016.

Kis Z, Shattock R, Shah N, Kontoravdi C. 2018. Emerging technologies for low-cost, rapid vaccine manufacture. *Biotechnology Journal* **14(1)**:e1800376 DOI 10.1002/biot.201800376.

Lafita A, Bliven S, Prlić A, Guzenko D, Rose PW, Bradley A, Pavan P, Myers-Turnbull D, Valasatava Y, Heuer M, Larson M, Burley SK, Duarte JM. 2019. BioJava 5: a

community driven open-source bioinformatics library. *PLOS Computational Biology* **15**:e1006791 DOI 10.1371/journal.pcbi.1006791.

Lobry JR, Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research* **22**:3174–3180 DOI 10.1093/nar/22.15.3174.

Merkel D. 2014. Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal* **2014**:2.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Research* **49**:D412–D419 DOI 10.1093/nar/gkaa913.

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. 2021. Sustainable data analysis with Snakemake. *F1000Research* **10**:33 DOI 10.12688/f1000research.29032.2.

Möller S, Croning MDR, Apweiler R. 2001. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**:646–653 DOI 10.1093/bioinformatics/17.7.646.

Pierleoni A, Martelli PL, Casadio R. 2008. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* **9**:392 DOI 10.1186/1471-2105-9-392.

Prianichnikov N, Koch H, Koch S, Lubeck M, Heilig R, Brehmer S, Fischer R, Cox J. 2020. MaxQuant software for ion mobility enhanced shotgun proteomics. *Molecular & Cellular Proteomics* **19**:1058–1069 DOI 10.1074/mcp.TIR119.001720.

Sigrist CJA, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. 2013. New and continuing developments at PROSITE. *Nucleic Acids Research* **41**:D344–D347 DOI 10.1093/nar/gks1067.

Springhorn A, Hoppe T. 2019. Western blot analysis of the autophagosomal membrane protein LGG-1/LC3 in *Caenorhabditis elegans*. In: *Methods in enzymology, vol 619*. Elsevier, 319–336 DOI 10.1016/bs.mie.2018.12.034.

Tamura K, Stecher G, Kumar S. 2021. MEGA11: molecular evolutionary genetics analysis version 11. *Molecular Biology and Evolution* **38**:3022–3027 DOI 10.1093/molbev/msab120.

The Galaxy Community, Abueg LAL, Afgan E, Allart O, Awan AH, Bacon WA, Baker D, Bassetti M, Batut B, Bernt M, Blankenberg D, Bombarely A, Bretaudeau A, Bromhead CJ, Burke ML, Capon PK, Čech M, Chavero-Díez M, Chilton JM, Collins TJ, Coppens F, Coraor N, Cuccuru G, Cumbo F, Davis J, De Geest PF, De Koning W, Demko M, DeSanto A, Begines JMD, Doyle MA, Droesbeke B, Erxleben-Eggenhofer A, Föll MC, Formenti G, Fouilloux A, Gangazhe R, Genthon T, Goecks J, Beltran ANG, Goonasekera NA, Goué N, Griffin TJ, Grüning BA, Guerler A, Gundersen S, Gustafsson OJR, Hall C, Harrop TW, Hecht H, Heidari A, Heisner T, Heyl F, Hiltemann S, Hotz H-R, Hyde CJ, Jagtap PD, Jakiela J, Johnson JE, Joshi J, Jossé M, Jum'ah K, Kalaš M, Kamieniecka K, Kayikcioglu T,

Konkol M, Kostrykin L, Kucher N, Kumar A, Kuntz M, Lariviere D, Lazarus R, Bras YL, Corguillé GL, Lee J, Leo S, Liborio L, Libouban R, Tabernero DL, Lopez-Delisle L, Los LS, Mahmoud A, Makunin I, Marin P, Mehta S, Mok W, Moreno PA, Morier-Genoud F, Mosher S, Müller T, Nasr E, Nekrutenko A, Nelson TM, Oba AJ, Ostrovsky A, Polunina PV, Poterlowicz K, Price EJ, Price GR, Rasche H, Raubenolt B, Royaux C, Sargent L, Savage MT, Savchenko V, Savchenko D, Schatz MC, Seguineau P, Serrano-Solano B, Soranzo N, Srikakulam SK, Suderman K, Syme AE, Tangaro MA, Tedds JA, Tekman M, Cheng (Mike) Thang W, Thanki AS, Uhl M, Van Den Beek M, Varshney D, Vessio J, Videm P, Von Kuster G, Watson GR, Whitaker-Allen N, Winter U, Wolstencroft M, Zambelli F, Zierep P, Zoabi R. 2024. The galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research* **52**:W83–W94 DOI 10.1093/nar/gkae410.

The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bursteinas B, Bye-A-Jee H, Coetzee R, Cukura A, Da Silva A, Denny P, Dogan T, Ebenezer T, Fan J, Castro LG, Garmiri P, Georghiou G, Gonzales L, Hatton-Ellis E, Hussein A, Ignatchenko A, Insana G, Ishtiaq R, Jokinen P, Joshi V, Jyothi D, Lock A, Lopez R, Luciani A, Luo J, Lussi Y, MacDougall A, Madeira F, Mahmoudy M, Menchi M, Mishra A, Moulang K, Nightingale A, Oliveira CS, Pundir S, Qi G, Raj S, Rice D, Lopez MR, Saidi R, Sampson J, Sawford T, Speretta E, Turner E, Tyagi N, Vasudev P, Volynkin V, Warner K, Watkins X, Zaru R, Zellner H, Bridge A, Poux S, Redaschi N, Aimo L, Argoud-Puy G, Auchincloss A, Axelsen K, Bansal P, Baratin D, Blatter M-C, Bolleman J, Boutet E, Breuza L, Casals-Casas C, De Castro E, Echioukh KC, Coudert E, Cuche B, Doche M, Dornevil D, Estreicher A, Famiglietti ML, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Hyka-Nouspikel N, Jungo F, Keller G, Kerhornou A, Lara V, Mercier PLe, Lieberherr D, Lombardot T, Martin X, Masson P, Morgat A, Neto TB, Paesano S, Pedruzzi I, Pilbout S, Pourcel L, Pozzato M, Pruess M, Rivoire C, Sigrist C, Sonesson K, Stutz A, Sundaram S, Tognolli M, Verbregue L, Wu CH, Arighi CN, Arminski L, Chen C, Chen Y, Garavelli JS, Huang H, Laiho K, McGarvey P, Natale DA, Ross K, Vinayaka CR, Wang Q, Wang Y, Yeh L-S, Zhang J, Ruch P, Teodoro D. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**:D480–D489 DOI 10.1093/nar/gkaa1100.

Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, Mi H. 2022. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Science* **31**:8–22 DOI 10.1002/pro.4218.

Vaudel M, Verheggen K, Csordas A, Ræder H, Berven FS, Martens L, Vizcaíno JA, Barsnes H. 2016. Exploring the potential of public proteomics data. *Proteomics* **16**:214–225 DOI 10.1002/pmic.201500295.

Waskom M. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* **6**:3021 DOI 10.21105/joss.03021.

Wratten L, Wilm A, Göke J. 2021. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods* **18**:1161–1168 DOI 10.1038/s41592-021-01254-9.

Ye J, Zhang Y, Cui H, Liu J, Wu Y, Cheng Y, Xu H, Huang X, Li S, Zhou A, Zhang X, Bolund L, Chen Q, Wang J, Yang H, Fang L, Shi C. 2018. WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Research* **46**:W71–W75 DOI 10.1093/nar/gky400.