Published online 22 May 2021 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* ***W169–W173***
https://doi.org/10.1093/nar/gkab391

DomainViz: intuitive visualization of consensus domain distributions across groups of proteins

Pascal Schläpfer [1,†], Devang Mehta [2,†], Cameron Ridderikhoff[2] and R. Glen Uhrig [2,*]

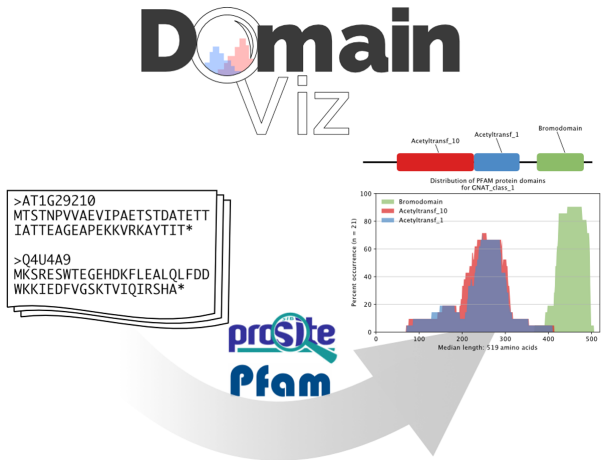
[1]Institute for Molecular Plant Biology, D-BIOL, ETH Zurich, Zürich 8092, Switzerland and [2]Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada

Received March 08, 2021; Revised April 14, 2021; Editorial Decision April 27, 2021; Accepted April 28, 2021

ABSTRACT

The prediction of functional domains is typically among the first steps towards understanding the function of new proteins and protein families. There are numerous databases of annotated protein domains that permit researchers to identify domains on individual proteins of interest. However, it is necessary to perform high-throughput domain searches to gain evolutionary insight into the functions of proteins and protein families. Unfortunately, at present, it is difficult to search for, and visualize domain conservation across multiple proteins and/or multiple groups of proteins in an intuitive manner. Here we present DomainViz, a new web-server that streamlines the identification and visualization of domains across multiple protein sequences. Currently, DomainViz uses the well-established PFAM and Prosite databases for domain searching and assembles intuitive, publication-ready 'monument valley' plots (mv-plots) that display the extent of domain conservation along two dimensions: positionality and frequency of occurrence in the input protein sequences. In addition, DomainViz produces a conventional domain-ordering figure. DomainViz can be used to explore the conservation of domains within a single protein family, across multiple families, and across families from different species to support studies into protein function and evolution. The web-server is publicly available at: https://uhrigprotools.biology.ualberta.ca/domainviz.

GRAPHICAL ABSTRACT



INTRODUCTION

Predicting the presence of domains is a fundamental step in assigning function to new proteins and protein families. The identification of conserved protein sequences and domains has been a long-standing endeavour in bioinformatics and has resulted in the creation of databases like PFAM, Prosite and InterPro that provide critical information for genomics projects (1–3). The classification of protein domains is typically based on the identification of sequence patterns based on evolutionary conservation, the presence of structural features, and / or their functional associations. As a result, comparing the presence and architecture of protein domains across entire protein families, or across evolutionary lineages, can provide valuable insights into the evolution of protein function by elucidating domain gain, loss or re-arrangement.

Indeed, comparative domain structure analysis has become commonplace in protein evolution studies (4–7). However, such an analysis invariably involves reducing the complexity of domain presence and ordering within protein sequences to binary representations of domain presence or absence (Figure 1A). Such representations elide impor-

*To whom correspondence should be addressed. Tel: +1 780 492 3088; Email: ruhrig@ualberta.ca
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

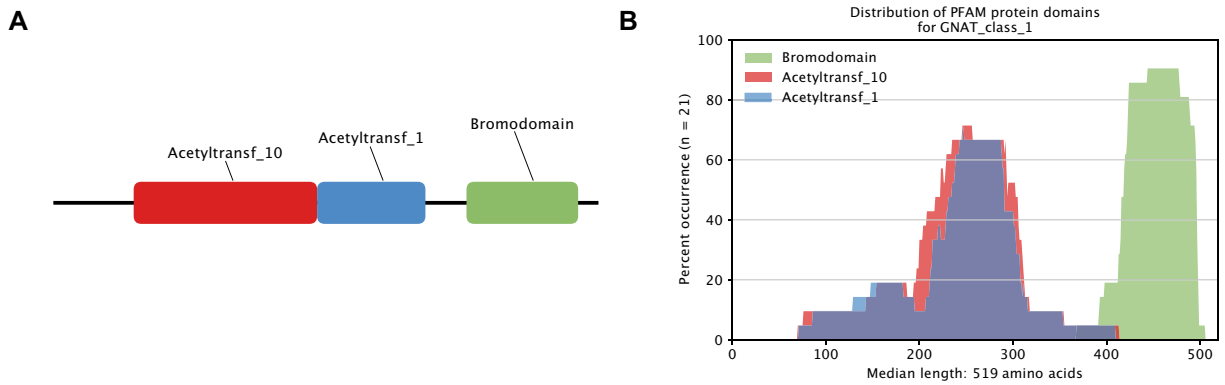


Figure 1. Protein domain representation of the Class 1 GCN5/PCAF-related N-acetyltransferases (GNATs). (**A**) Conventional representation of PFAM domain order within a group of 21 GNAT proteins. (**B**) mv-plot representation of PFAM domain conservation within the same group of proteins. Both images have been generated using DomainViz.

tant details such as the relative distribution of protein domains within evolutionary clades and the degree to which domain sizes are conserved within protein groups and families. In 2017, we pioneered a new format for displaying protein domain structure in an evolutionary analysis of regulatory protein acetylation enzymes (Figure 1B) (8). This new quantitative representation depicted domain architecture in two dimensions, with the position and size of protein domains on the horizontal axis, and conservation frequency on the vertical axis. This representation that we dub 'mv-plots' (due to their visual similarity to the sandstone buttes in Arizona's Monument Valley) is information rich and provides viewers with an intuitive grasp of domain presence, position and conservation.

Constructing mv-plots manually is however, a long and laborious process involving several steps. It involves multiple queries of domain databases (PFAM, Prosite, etc.), calculating domain conservation and position, and then plotting these data points, and then repeating this process for each domain found in the initial query set. Here, we present DomainViz, a web-server based tool that can generate mv-plots of domain conservation for multiple protein sequences, and multiple groups of protein sequences in a single step. DomainViz offers distinct functionality compared to other domain visualization tools such as SMART (9) and POLYVIEW (10). Most notably, DomainViz creates two-dimensional outputs that summarize domain conservation across multiple proteins whereas SMART and POLYVIEW output domain position information on single proteins only. However, DomainViz does not predict structural features, and POLYVIEW is more appropriate for such investigation. SMART is also more appropriate for domain specific investigations such as studies that wish to search for all proteins that share a specific user-defined domain. Thus, DomainViz adds to this core set of protein domain analysis tools and fills an important niche in conducting domain analysis across multiple proteins, which is particularly relevant for evolutionary and meta-analysis studies.

DomainViz uses both PFAM and Prosite databases and simply requires a multi-sequence FASTA file of protein sequences as input to generate publication quality mv-plots. In addition, DomainViz also provides a more conventional domain-ordered protein representation that precisely maps each domain's location based on mv-plot data. The outputs, (including intermediate PFAM and Prosite outputs) are downloadable, and editable with any vector image processing software.

METHODS AND RESULTS

Implementation

The webserver was created using nginx (https://www.nginx.com/) to serve the frontend, with a reverse proxy to the backend, which is served using gunicorn (https://gunicorn.org/) by following the guide created by Miguel Grinberg (11). This was then updated to serve onto HTTPS via a second Grinberg' guide (12). The outward facing website was designed using MaterialUI and the server is set up on a remote virtual Linux machine (Ubuntu 20.04.1 LTS) hosted on the University of Alberta Cirrus Compute and Storage Cloud (https://spaces.facsci.ualberta.ca/cirrus/). The backend consists of a single script, domainviz.py, containing all functions and modules to process input sequences. The script was developed with Python 3.8.2 and has been tested on Linux and macOS. The program uses python modules *sys*, *os*, *pathlib*, *shutil*, *distutils* to handle files, folders, paths and job abortion. We use *re* to perform string manipulations, and numpy and pandas to read, store, and manipulate arrays of data. *math*, *numpy*, and *statistics* are employed for statistical calculations. We use *matplotlib*, *mpl_toolkits* and *plotly* to plot and save images. *Bio* was used to handle Prosite records and *urllib* to retrieve PFAM records. We use *time* and *datetime* to measure time.

Usage

The DomainViz home-page features a central input panel where users can upload one or more protein FASTA files. Once the file(s) is uploaded, users have the option to rename files with the display name to be used in the final results page. Multiple files can either be uploaded at once, or sequentially. Users can then proceed to click the 'Submit' button to launch the job with default parameters, or choose to modify the default parameters. These parameters are:

1. Minimum domain prevalence: Only domains that appear more frequently in each input file than the defined value (default 0.05, or 5%), are displayed.
2. Minimum domain position conservation: Only domains that are present more frequently than the defined value (default 0.05, or 5%) at a single position on the consensus protein are displayed.
3. Figure scaling: This optional parameter scales the size of the image output to the median length of the input protein sequences. (We recommend first-time users operate DomainViz with this option deactivated.)
4. Absolute y-axis: If selected, the output image will display the conservation frequency in absolute numbers of proteins, rather than percentage of input proteins.

Upon job submission, a new processing page is loaded where the unique result id is displayed. This result id can be saved to retrieve results at a later time using the 'Result Id' text field on the home-page. The processing page will reload showing the progress of the task until completion.

Input processing

The DomainViz computation pipeline is summarized in Figure 2, beginning with users uploading multi-sequence FASTA files. In the event that a user uploads more than one FASTA input, all files are compiled into a single FASTA file with sequence headers amended to include the user defined protein group name, or if absent the source file name. Subsequently, all proteins are queried against an internal database of PFAM and Prosite results to find pre-processed data. Proteins that do not have existing results in our internal database are then queried against PFAM and Prosite. Prosite is queried via biopython (Bio.ExPASy.ScanProsite) accessing https://prosite.expasy.org. PFAM domains are queried via a custom script that accesses the hmmer website (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan) and searches for PFAM domains.

New results are then added to the internal database to accelerate future queries. The internal database only stores domain prediction results, the amino acid sequence of the query protein, and the date of result retrieval. No other information such as the primary job id, IP address, or even the FASTA header of the query sequence is stored. The internal database is structured as a set of files named by the first five amino acids of the constituent protein sequences. Results of proteins with the same five first amino acids are hence grouped, speeding up data retrieval.

Once the protein domain prediction results are retrieved, a consensus visualization of the results for all proteins in each input file needs to be generated. Since proteins in a single input group are typically not of identical length, each group of proteins has its own distribution of lengths that must be accounted for. After testing various measures (maximal size, median, mean, minimal size, most prevalent size), we decided that results are most easily understandable when the median protein length of the group is used on one axis. Hence, the median length of the protein group is computed. Next, the prevalence of each domain across the median length is calculated to produce a histogram. The prevalence of each domain per amino acid of the median length protein is computed by:

$$\rho_{a,D,P} = \frac{\sum_{i=1}^{n_P} 1_{F_i}}{n_P}$$

$$where\ F = \left(\frac{S_{i,D} \cdot m_P}{l_i} \leq a \leq \frac{E_{i,D} \cdot m_P}{l_i} \right)$$

$\rho_{a,D,P}$ is the prevalence ρ of domain D at position a of the hypothetical median length of protein group P, n_P is the number of proteins in the protein group, $S_{i,D}$ is the start of the domain D in the ith protein, m_P is the median amino acid length of the protein group, l_i is the length of the ith protein, and $E_{i,D}$ is the end of the domain D in the ith protein. Histograms are then produced for each domain and plotted together to make a single mv-plot (Figure 2).

RESULTS

Upon job completion, the progress page redirects to a Results page where three mv-plots for each protein group (one for PFAM, Prosite, and combined results, each) are displayed. Conventional domain representations are also presented adjacent to the respective mv-plot. These web-based results are interactive plots, where users can choose to view or hide selected domains using the interactive legend alongside the mv-plot. This is particularly useful when multiple overlapping domains are present. These interactive plots also possess a snapshot function, whereby users can quickly download their outputs. However, a downloadable compressed output file is also made available to users. The downloadable results include vector editable PDF files for each mv-plot, as well as tab separated text files for the PFAM and Prosite outputs, and for the calculated domain prevalence per amino acid. The Results page can be accessed after job submission by entering the unique job id in the relevant field on the home page.

Processing times

The processing time of a DomainViz query can vary considerably depending on the type and number of protein sequences entered. Results are displayed almost instantaneously (within 30 s) for queries exclusively containing proteins that have previously been queried using DomainViz. As a result, over time, processing times for most common sequences can be expected to be relatively short. The median processing time per sequence for completely new queries is 21 s and so very large queries consisting of thousands of sequences can even take hours to process. Hence, we recommend users save the unique result id generated per query in order to check for results at later times. Additionally, the processing page for each query automatically updates with the status of the job allowing users to estimate the total time for completion.

Application scenario(s)

DomainViz has built in example files that users can load to test the program and parameters. These examples are derived from our 2017 genome-scale analysis of enzymes involved in regulatory lysine acetylation (8). This analysis is

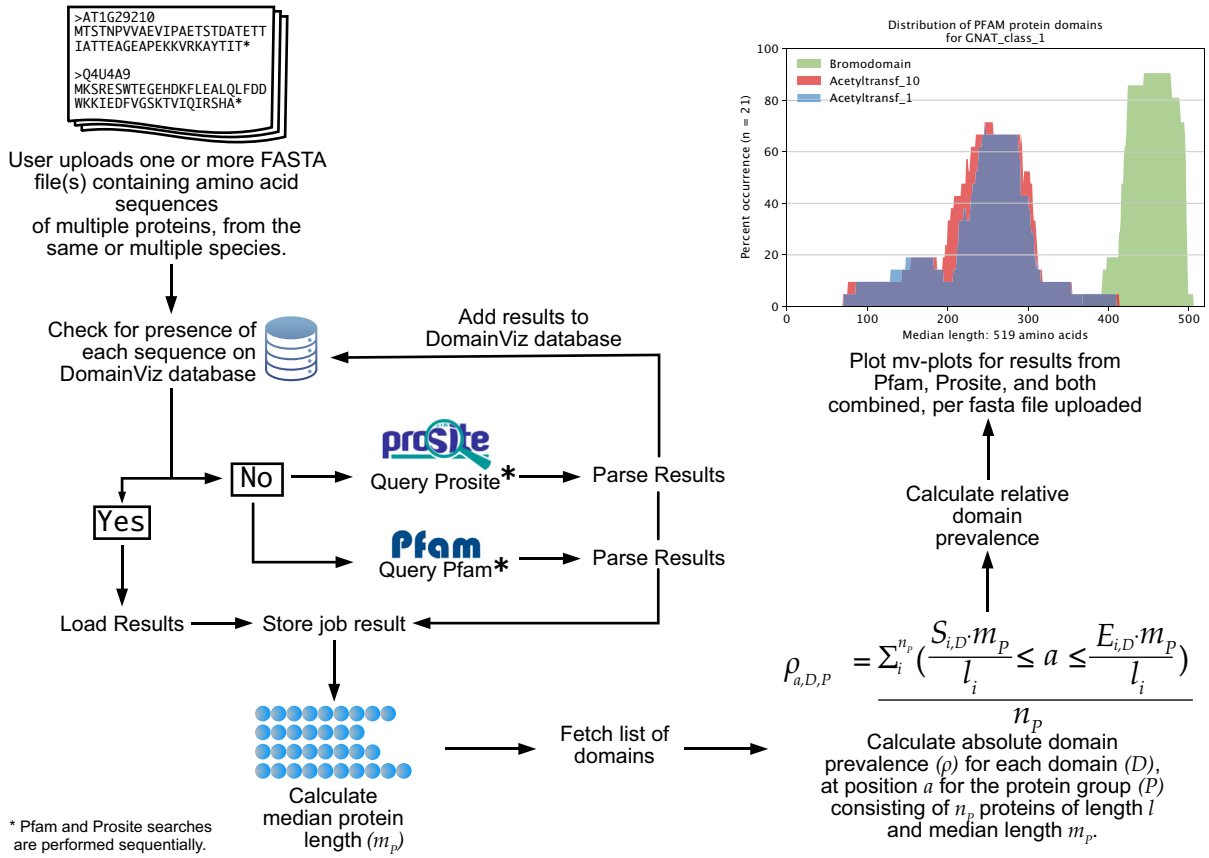


Figure 2. Schematic representation of the DomainViz computational pipeline.

an example of a typical application scenario involving the domain conservation analysis of a large family of proteins across a wide range of species.

N^{ε}-Lysine acetylation has recently emerged as a wide-ranging regulatory post-translational modification in eukaryotes (8,13,14). As a result, characterization of N^{ε}-lysine acetylation catalyzing enzymes (N^{ε}-lysine acetyltransferases and deacetylases) from across photosynthetic eukaryotes was essential. First, consensus phylogenetic trees were derived for each N^{ε}-lysine acetyltransferase and deacetylase family using both maximum likelihood and bayesian analyses (8). For the GCN5/PCAF-related N-acetyltransferases (GNATs) this uncovered three phylogenetically distinct subclasses of GNAT proteins. Upon subjecting a selection of sequences from these three subclasses to DomainViz analysis, we found that each GNAT class had distinct domain composition, positionality and abundance, providing functional context to the initial phylogenetic analysis. In particular, sub-class I GNATs maintained a highly conserved acetyltransferase domain and c-terminal bromodomain structure, while sub-class II possessed a conserved HAT1-N acetyltransferase domain. In addition, 60% of sub-class II GNAT sequences uniquely possessed a second acetyltransferase. GNAT sub-class III proteins, on the other hand, possessed a C-terminal acetyltransferase domain with an N-terminal radical-SAM domain, which are implicated in catalyzing an array of unique reactions ranging from post-transcriptional and post-translational modifications to aspects of lipid metabolism and co-factor biosynthesis (15). Thus, in this use case a DomainViz-style analysis provided significant added value by elucidating the distribution, positionality and conservation of each domain across the three subclasses of the photosynthetic eukaryote GNAT acetyltransferase family, which can now be used to guide future experimentation.

DomainViz now makes such analyses available to all researchers interested in protein evolution by offering an easy-to-use web service that produces intuitive, easy to understand, high resolution visualizations to drive discovery.

Planned features

We plan to continually improve the performance and feature set offered by DomainViz. One of our planned features in the short term is background querying of random protein sequences from GenBank against PFAM and Prosite in order to regularly populate and re-populate our internal database in order to speed up job processing. This will also allow us to more accurately predict estimated times for job completion in order to improve the user experience. Further, we will allow users to select their desired colors to be used in constructing mv-plots in order to eliminate any need for additional image editing. A major planned update will also enable users to automatically delineate groups of proteins

by clustering input sequences based on the calculated domain prevalences, thus directly supporting the generation of new evolutionary hypotheses.

DISCUSSION

Existing protein domain databases like PFAM and Prosite possess information for tens of thousands of protein features, with the on-going revolution in genome sequencing continuing to add to this accumulated knowledge. However, currently, this data can only be accessed through extensive, time-consuming manual searching, and while frequently used in studies of protein family evolution, is rarely represented quantitatively. DomainViz offers scientists a single easy to use tool to query hundreds of protein sequences to retrieve domain data and generate a single quantitatively accurate representation of domain conservation for various domains across multiple proteins. We believe that DomainViz will become an essential tool in the analysis of new protein families and classes and will empower biological researchers of all stripes to gain an intuitive understanding of protein domain conservation.

DATA AVAILABILITY

DomainViz is available at https://uhrigprotools.biology.ualberta.ca/domainviz. Source code and data is available under a GNU Affero General Public License 3.0 at https://github.com/UhrigLab/DomainViz.

ACKNOWLEDGEMENTS

The authors thank John Bartoszewski and Broderick Wood of the Faculty of Science Research IT team (University of Alberta) for their assistance in setting up the web server, Mohamad Jamaleddine (University of Alberta) for troubleshooting, and Sukalp Muzumdar (Cold Spring Harbor Laboratory) for assistance with mathematical formulation.

FUNDING

Natural Sciences and Engineering Research Council (NSERC) of Canada. Fundig for open access charge: Natural Sciences and Engineering Research Council of Canada
Conflict of interest statement. None declared.

REFERENCES

1. Sigrist,C.J.A., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinformatics*, **3**, 265–274.
2. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
3. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
4. Saito,M., Sato,A., Nagata,S., Tamaki,S., Tomita,M., Suzuki,H. and Kanai,A. (2019) Large-scale molecular evolutionary analysis uncovers a variety of polynucleotide kinase Clp1 family proteins in the three domains of life. *Genome Biol. Evol.*, **11**, 2713–2726.
5. Mutte,S.K. and Weijers,D. (2020) Deep evolutionary history of the phox and bem1 (PB1) domain across eukaryotes. *Sci. Rep.*, **10**, 3797.
6. Dievart,A., Gottin,C., Périn,C., Ranwez,V. and Chantret,N. (2020) Origin and diversity of plant receptor-like kinases. *Annu. Rev. Plant Biol.*, **71**, 131–156.
7. Liu,B.A. and Nash,P.D. (2012) Evolution of SH2 domains and phosphotyrosine signalling networks. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.*, **367**, 2556–2573.
8. Uhrig,R.G., Schläpfer,P., Mehta,D., Hirsch-Hoffmann,M. and Gruissem,W. (2017) Genome-scale analysis of regulatory protein acetylation enzymes from photosynthetic eukaryotes. *BMC Genomics*, **18**, 514.
9. Letunic,I. and Bork,P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.*, **46**, D493–D496.
10. Porollo,A.A., Adamczak,R. and Meller,J. (2004) POLYVIEW: a flexible visualization tool for structural and functional annotations of proteins. *Bioinformatics*, **20**, 2460–2462
11. Grinberg,M. (2020) In: *How to Deploy a React + Flask Project*. https://blog.miguelgrinberg.com/post/how-to-deploy-a-react--flask-project.
12. Grinberg,M. (2017) In: *Running Your Flask Application Over HTTPS*. https://blog.miguelgrinberg.com/post/running-your-flask-application-over-https.
13. Hartl,M., Füßl,M., Boersema,P.J., Jost,J.-O., Kramer,K., Bakirbas,A., Sindlinger,J., Plöchinger,M., Leister,D., Uhrig,G. *et al.* (2017) Lysine acetylome profiling uncovers novel histone deacetylase substrate proteins in Arabidopsis. *Mol. Syst. Biol.*, **13**, 949.
14. Uhrig,R.G., Schläpfer,P., Roschitzki,B., Hirsch-Hoffmann,M. and Gruissem,W. (2019) Diurnal changes in concerted plant protein phosphorylation and acetylation in Arabidopsis organs and seedlings. *Plant J.*, **99**, 176–194.
15. Holliday,G.L., Akiva,E., Meng,E.C., Brown,S.D., Calhoun,S., Pieper,U., Sali,A., Booker,S.J. and Babbitt,P.C. (2018) Atlas of the radical SAM superfamily: divergent evolution of function using a "plug and play" domain. *Methods Enzymol.*, **606**, 1–71.