PDB ProtVista: A reusable and open-source sequence feature viewer

Mandar Deshpande[1], Mihaly Varadi[1*], Typhaine Paysan-Lafosse[1], Sreenath Nair[1], Damiano Piovesan[2], Saqib Mir[1], Aleksandras Gutmanas[1], Silvio C. E. Tosatto[2], Sameer Velankar[1]

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK

2 University of Padua, Department of Biomedical Sciences, Padova, Italy

* Corresponding author

Abstract

Summary

The PDB ProtVista is a reusable and customisable sequence feature viewer that provides intuitive and detailed 2D visualisation of residue-level annotations while also supporting interactive communication with 3D viewers. The Protein Data Bank in Europe (PDBe) team develops and maintains PDB ProtVista. Several web services use it to display structural and functional annotations such as macromolecular interaction interfaces, intrinsic disorder predictions, sequence variants, and sequence conservation.

Availability and Implementation

The PDB ProtVista is a native JavaScript (ES6) application that is freely available from https://github.com/PDBeurope/protvista-pdb. The above URL also provides extensive documentation and step-by-step user guides on integrating PDB ProtVista with existing web applications and 3D molecular viewers. It also offers an example of how to display custom data of the user and functional annotations for PDB and UniProt entries powered by a rich set of PDBe API endpoints.

1 Introduction

Identifying the biological function of macromolecules is instrumental in understanding and potentially modulating cellular processes (Waman et al., 2021). Structural representations of proteins and nucleic acids help researchers unveil these mechanisms (D'Amico et al., 2021; Ravarani et al., 2020). A wealth of structure- and sequence-derived functional and biophysical annotations can further enhance the inherent value of macromolecular models (wwPDB consortium, 2018; Varadi et al., 2019).

In 2018, we have established the Protein Data Bank in Europe - Knowledge Base (PDBe-KB) that integrates macromolecular structural data with annotations provided by data resources and research teams from across the globe (Varadi et al., 2019). However, the complexity of these data makes them challenging to interpret and gain biological insights unless advanced and yet easy-to-use data visualisation tools are readily available to the scientific community (Sehnal et al., 2020). A powerful and intuitive way of visualising annotations is to use so-called sequence feature viewers to overview all the annotated amino acid or nucleic acid residues for a particular macromolecular sequence (Watkins et al., 2017).

Here, we present PDB ProtVista, a sequence feature viewer with native support for interactions with 3D molecular graphics viewers, that utilises and expands the Nightingale library (https://github.com/ebi-webcomponents/nightingale), which is a collection of web components co-developed by PDBe (Armstrong et al., 2019), UniProt (Bateman et al., 2021) and InterPro (Blum et al., 2021). PDBe and PDBe-KB web pages use the PDB ProtVista to visualise annotations and map them interactively onto the 3D models. PDB ProtVista is open-source and available from https://github.com/PDBeurope/protvista-pdb.

2 Materials and methods

Based on the first version of ProtVista (Watkins et al., 2017), the PDBe, UniProt, and InterPro teams created a collection of web components called the Nightingale library (https://github.com/ebi-webcomponents/nightingale). PDB ProtVista extends and tailors these web components to create an implementation specifically for displaying structural annotations found at PDBe and PDBe-KB. PDB ProtVista is an easily reusable, framework-agnostic sequence feature viewer that seamlessly communicates with the 3D-molecular graphics viewer, Mol* (Sehnal et al., 2021).

The PDB ProtVista provides three data visualisation modes, called “tracks”, to support various types of sequence and structure annotations: i.) a “segment” track; ii.) a “variants” track, and iii.) a “sequence logo” track. The “segment” track, the most frequently used visualisation mode, displays block representations of either a single position in a sequence or a continuous range of residues. PDBe and PDBe-KB use this visualisation mode to display annotations such as structural domains (Andreeva et al., 2020; Sillitoe et al., 2019), consensus disorder predictions from the MobiDB data resource (Piovesan et al., 2021) or continuous variables such as backbone flexibility propensity (Cilia et al., 2014).

Amino acid variants can be displayed using a dedicated data visualisation track that supports displaying distinct sequence positions and their corresponding meta-information, such as the biological effect of mutations. PDBe and PDBe-KB use this track to display variants data from UniProt and sequence variants and their predicted consequences provided by Missense3D (Khanna et al., 2021) and FoldX (Delgado et al., 2019).

Finally, PDB ProtVista can show sequence conservation as a histogram together with the probability that a particular amino acid appears at a given position using a sequence logo

track developed using the D3.js library. ProtVista displays the sequence logo when the user zooms in on the shown sequence, and 150 or fewer residues are visible. We set this threshold to improve the loading performance and the user visualisation experience. Each amino acid is coloured using its electrochemical properties as defined by the Clustal X colour scheme (Larkin et al., 2007). For each position, PDB ProtVista sorts the amino acids according to their properties by default. Users may also sort the sequence logo by probability using the corresponding option on the left-hand side of the track.

The PDBe weekly release process includes these sequence conservation calculations, ensuring up-to-date data are displayed. The weekly process generates sequence conservation information for every PDB entry and corresponding UniProt accessions when sufficient sequence data is available in UniProt. The protein sequence is aligned against the UniProt database to obtain a multiple sequence alignment, and the process generates a Hidden-Markov Model profile using the HMMER suite (Potter et al., 2018). The data pipeline then calculates amino acid probabilities and conservation scores from the profile obtained. Users can download the multiple sequence alignment (MSA) by clicking on the "Download MSA" button in PDB ProtVista.

An interactive demo that showcases how to display custom data provided by the user is available at https://embed.plnkr.co/plunk/nPNvqHhyamzHGrCT. Comprehensive documentation, user guides and further examples of displaying data provided by the PDBe Aggregated API endpoints and linking PDB ProtVista to the 3D-viewer Mol* are publicly available from the PDB ProtVista GitHub repository (https://github.com/PDBeurope/protvista-pdb/wiki).

3 Discussion and conclusion

PDBe-KB integrates macromolecular structure data with functional and biophysical annotations in a comprehensive knowledge graph (Varadi et al., 2019). The data in this archive is complex and diverse, necessitating advanced and easy-to-use data visualisations to help the scientific community gain valuable insights regarding the biological role of macromolecules.

The PDB ProtVista is a sequence feature viewer consisting of reusable, framework-agnostic web components that can provide an overview of structural and sequence annotations while also supporting seamless communication with the 3D-molecular graphics viewer Mol* (Sehnal et al., 2021).

We showcase the power of PDB ProtVista through the example of Calpastatin (UniProt accession P27321), an inhibitor of the calcium-dependent, non-lysosomal cysteine protease Calpain (Figure 1).

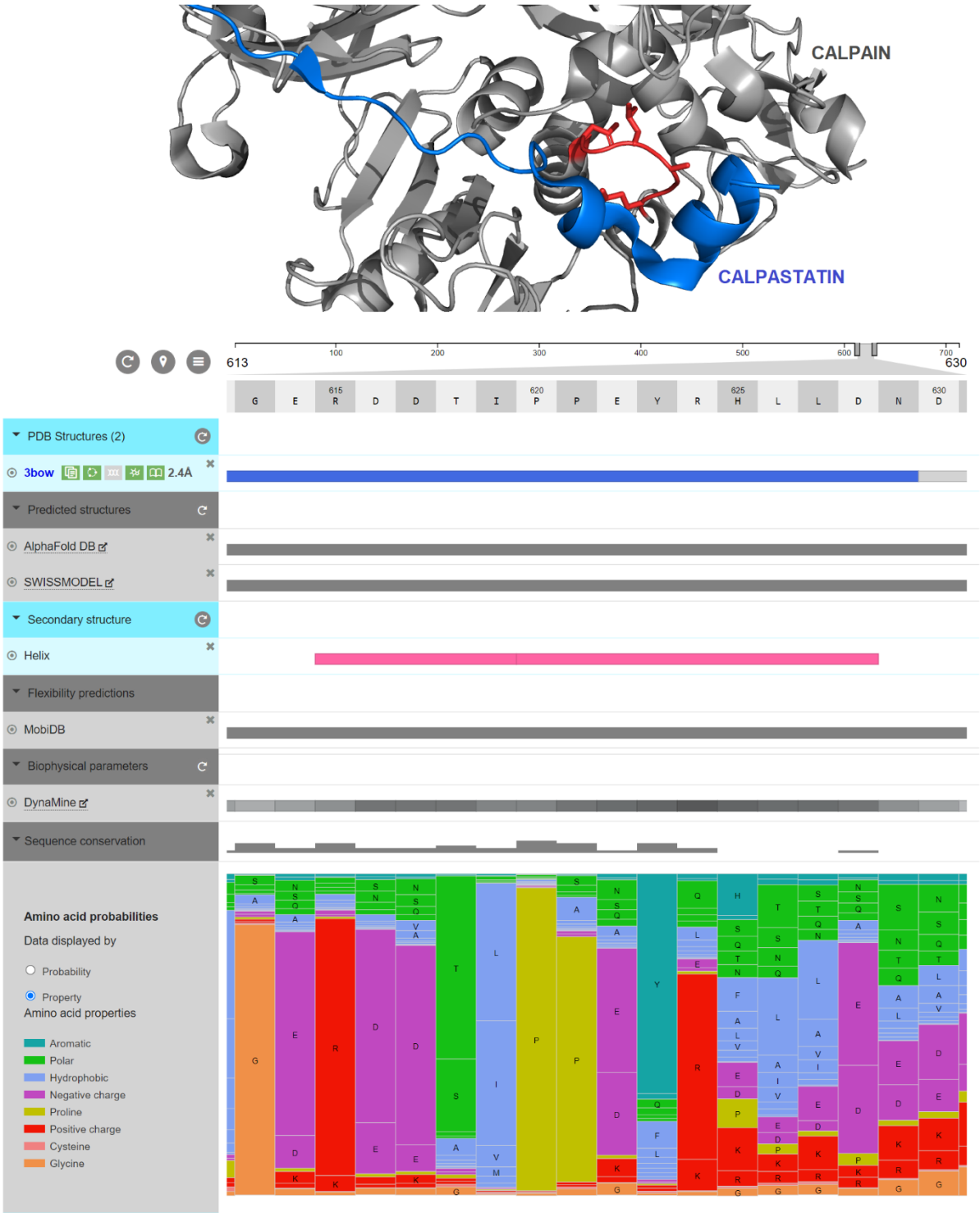

Figure 1 - PDB ProtVista is displaying annotations for Calpastatin.

Calpastatin is the native inhibitor of the protease Calpain. Using publicly available data, PDB ProtVista can display all the related PDB entries and predicted models from AlphaFold DB and SWISS-MODEL. It can also show the observed secondary structure elements (where the pink rectangle represents a helix) and all the

amino acid positions that interact with Calpain (blue rectangles). PDB ProtVista can also show computationally derived annotations (in grey), such as consensus disorder predictions from MobiDB, backbone flexibility propensities by DynaMine and sequence conservation propensities complemented by a sequence logo. Examining these annotations highlights the region where Calpastatin adopts a helical structure. This region bends out over the catalytic site of Calpain (coloured red on the figure above) to avoid enzymatic cleavage. These PDB ProtVista visualisation tracks are used at https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P27321 to display Calpastatin annotations.

Using publicly available API endpoints (Nair et al., 2021), PDB ProtVista can list all the known structures of Calpastatin in the PDB and predicted models from AlphaFold DB and SWISS-MODEL (Tunyasuvunakool et al., 2021; Waterhouse et al., 2018). It also displays information on the observed secondary structure elements, complemented by consensus disorder predictions from MobiDB and backbone flexibility predictions from DynaMine. PDB ProtVista can also highlight the interaction interfaces between Calpastatin and Calpain and display sequence conservation and sequence logo data. Together, these data visualisation tracks show that Calpastatin is an intrinsically disordered protein, according to MobiDB, and it has regions that may adopt context-dependent secondary structure, according to DynaMine. These regions adopt short helices in the Calpain/Calpastatin complex (PDB structure: 3bow) (Hanna et al., 2008). PDB ProtVista also highlights that the sequence conservation propensity is higher for residues that do not directly interact with Calpain. This seemingly counterintuitive observation reflects the actual mechanism by which Calpastatin inhibits Calpain. The inhibitor protein binds to a long cleft on the catalytic subunit of Calpain. To avoid being cleaved by Calpain, Calpastatin loops out by forming a short helical secondary structure element, explaining why the sequence conservation of this region is higher than that of the flanking macromolecular interaction interfaces.

In addition to providing such functional insights through the 2D overview of annotations, an essential feature of PDB ProtVista is its capacity for seamless interaction with the 3D molecular graphics viewer, Mol* (Sehnal et al., 2021). For example, users can compare intrinsic disorder predictions with the observed interaction interfaces and secondary structure elements in 2D and 3D. The MobiDB data resource provides these up-to-date consensus disorder annotations that PDBe-KB integrates weekly with the core PDB data (Varadi et al., 2019). Two-way communication between Mol* and PDB ProtVista makes it easy to interactively identify regions that are likely to adapt semi-stable secondary structures when bound to their interaction partners, such as Calpastatin's case described above.

In conclusion, PDB ProtVista is a robust, reusable and intuitive sequence feature viewer that specialises in displaying structural and functional annotations in 2D, with built-in support for mapping these annotations onto 3D molecular representations. It supports data provided by the users and data from a rich collection of API endpoints offered by PDBe.

Funding

We would like to acknowledge funding from ELIXIR [IDP implementation study], the Biotechnology and Biological Sciences Research Council via the 3D-Gateway [BB/T01959X/1], Genome3D (BB/N019172/1), and FunPDBe [BB/P024351/1] grants, the Wellcome Trust [104948] and the European Molecular Biology Laboratory-European Bioinformatics Institute who supported this work.

References

Andreeva, A., Kulesha, E., Gough, J., & Murzin, A. G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, *48*(D1). https://doi.org/10.1093/nar/gkz1064

Armstrong, D. R., Berrisford, J. M., Conroy, M. J., Gutmanas, A., Anyango, S., Choudhary, P., Clark, A. R., Dana, J. M., Deshpande, M., Dunlop, R., Gane, P., Gáborová, R., Gupta, D., Haslam, P., Koča, J., Mak, L., Mir, S., Mukhopadhyay, A., Nadzirin, N., … Velankar, S. (2019). PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz990

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L. G., … Teodoro, D. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa1100

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., … Finn, R. D. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa977

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., & Vranken, W. F. (2014). The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Research*, *42*(W1). https://doi.org/10.1093/nar/gku270

consortium, wwPDB. (2018). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky949

D'Amico, F., Candido, S., & Libra, M. (2021). Interaction between matrix metalloproteinase-9 (MMP-9) and neutrophil gelatinase-associated lipocalin (NGAL): A recent evolutionary event in primates. *Developmental & Comparative Immunology*, *116*. https://doi.org/10.1016/j.dci.2020.103933

Delgado, J., Radusky, L. G., Cianferoni, D., & Serrano, L. (2019). FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*, *35*(20). https://doi.org/10.1093/bioinformatics/btz184

Hanna, R. A., Campbell, R. L., & Davies, P. L. (2008). Calcium-bound structure of calpain and its mechanism of inhibition by calpastatin. *Nature*, *456*(7220). https://doi.org/10.1038/nature07451

Khanna, T., Hanna, G., Sternberg, M. J. E., & David, A. (2021). Missense3D-DB web catalogue: an atom-based analysis and repository of 4M human protein-coding genetic variants. *Human Genetics*, *140*(5). https://doi.org/10.1007/s00439-020-02246-z

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21). https://doi.org/10.1093/bioinformatics/btm404

Nair, S., Váradi, M., Nadzirin, N., Pravda, L., Anyango, S., Mir, S., Berrisford, J., Armstrong, D., Gutmanas, A., & Velankar, S. (2021). PDBe aggregated API: programmatic access to an integrative knowledge graph of molecular structure data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab424

Piovesan, D., Necci, M., Escobedo, N., Monzon, A. M., Hatos, A., Mičetić, I., Quaglia, F., Paladin, L., Ramasamy, P., Dosztányi, Z., Vranken, W. F., Davey, N. E., Parisi, G., Fuxreiter, M., & Tosatto, S.

C. E. (2021). MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa1058

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, *46*(W1). https://doi.org/10.1093/nar/gky448

Ravarani, C. N. J., Flock, T., Chavali, S., Anandapadamanaban, M., Babu, M. M., & Balaji, S. (2020). Molecular determinants underlying functional innovations of TBP and their impact on transcription initiation. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16182-z

Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S. K., Koča, J., & Rose, A. S. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkab314

Sehnal, D., Svobodová, R., Berka, K., Rose, A. S., Burley, S. K., Velankar, S., & Koča, J. (2020). High-performance macromolecular data delivery and visualization for the web. *Acta Crystallographica Section D Structural Biology*, *76*(12). https://doi.org/10.1107/S2059798320014515

Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., Tolulope, A., Scholes, H. M., Senatorov, I., Bujan, A., Ceballos Rodriguez-Conde, F., Dowling, B., Thornton, J., & Orengo, C. A. (2019). CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Research*, *47*(D1). https://doi.org/10.1093/nar/gky1097

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein

structure prediction for the human proteome. *Nature*. https://doi.org/10.1038/s41586-021-03828-1

Varadi, M., Berrisford, J., Deshpande, M., Nair, S. S., Gutmanas, A., Armstrong, D., Pravda, L., Al-Lazikani, B., Anyango, S., Barton, G. J., Berka, K., Blundell, T., Borkakoti, N., Dana, J., Das, S., Dey, S., Micco, P. di, Fraternali, F., Gibson, T., … Velankar, S. (2019). PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz853

Waman, V. P., Sen, N., Varadi, M., Daina, A., Wodak, S. J., Zoete, V., Velankar, S., & Orengo, C. (2021). The impact of structural bioinformatics tools and resources on SARS-CoV-2 research and therapeutic strategies. *Briefings in Bioinformatics*, *22*(2). https://doi.org/10.1093/bib/bbaa362

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*(W1). https://doi.org/10.1093/nar/gky427

Watkins, X., Garcia, L. J., Pundir, S., & Martin, M. J. (2017). ProtVista: visualization of protein sequence annotations. *Bioinformatics*, *33*(13). https://doi.org/10.1093/bioinformatics/btx120