





substrateminor: A Python package to investigate protein substrate repertoires

Robert QIAO ^{1,2}

¹ School of Biological Sciences, Flinders University, Bedford Park, SA 5042, Australia ² Digital Research Services, Flinders University, Bedford Park, SA 5042, Australia

DOI: [10.21105/joss.08266](https://doi.org/10.21105/joss.08266)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Charlotte Soneson](#) 

Reviewers:

- [@tushardave26](#)
- [@RiesBen](#)

Submitted: 26 February 2025

Published: 12 September 2025

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

Substrateminor is a Python library that provides a suite of modular investigative methods based on protein sequence and cellular properties to support rapid development of *in silico* pipelines to discover and interrogate potential substrates for a target protease of interest. Three categories of methods are included in substrateminor: consensus analysis, substrate mining, and pathological investigation. Together, they provide a well-suited toolkit to explore substrate repertoires, also known as the substrate degradome, for a target protease.

Statement of need

The ability to identify and discover novel protease *in situ* substrates is often a powerful first step towards understanding the biological functions of proteases and their implications in human health and disease. However, in traditional enzymology, substrates of a target enzyme are studied one at a time, arguably due to the prohibitive cost involved in designing and deploying large-scale experimental assays. Modern proteomics methods, particularly the advancement in mass spectrometry, drastically reduce the cost associated with detecting substrate hydrolysis events at the molecular level, and now the study of substrates of a target enzyme at the proteome level is no longer a distant fantasy. With these new possibilities, the lack of a modular system to enable rapid development of *in silico* pipelines to support preliminary discovery has been a bottleneck for new studies. Therefore, substrateminor aims to provide a versatile toolkit that enables fast modular development of *in silico* pipelines to facilitate the preliminary discovery that complements and accelerates the emerging experimental efforts.

Background

Proteolytic cleavage is a fundamental biochemical process in biology, which underpins many important biological processes that surround us from common food processing and digestion ([Campbell & Reece, 2005](#)) to protective immune responses against pathogen infections ([Kammerl & Meiners, 2016](#)). Almost all proteolytic processes in nature are catalysed by enzymes, and proteases are such an enzyme class that facilitates the biochemical reaction of cleaving peptide bonds. The inhibition of proteases has been a major intervention strategy in modern clinics to modulate molecular proteolytic processes to treat many physiological conditions including viral infections like HIV, hepatitis-C; metabolic dysfunctions like type-2 diabetes ([Scott, 2017](#)); and cancers ([Manasanch & Orlowski, 2017](#)). In the wake of the COVID-19 pandemic, proteases have emerged as a major therapeutic target for curbing fatality due to viral infections ([Borges et al., 2024](#); [Papaneophytou, 2024](#); [Sojka et al., 2021](#)).

Nevertheless, one constraining factor that confronts the rapid development of protease inhibitory therapy is the indiscriminating nature of inhibition. This is a major challenge because most

proteases are involved in a complex network of biological processes, and a complete inhibition of one protease can lead to unintended consequences on many other biological processes. Therefore, identifying and understanding the substrate degradome of a target protease is an important preliminary step towards inhibitory therapy. Traditionally, the identification of protease substrates is often challenging due to the prohibitive cost of large-scale deployment of target-designed immuno-based biochemistry assays and scarcity of adequate animal models. Modern advancements in proteomics, particularly the maturation of mass spectrometry (Han et al., 2008) and soft-ionisation methods (Challen & Cramer, 2022), started to become a cost-effective alternative in the identification of protease substrates at a proteome level. Therefore, the ability to rapidly develop *in silico* pipelines to support preliminary discovery of protease substrates is increasingly needed to complement the emerging experimental efforts.

Methods

Substrateminor contains three main classes of functions (with a schematic depicted in Figure 1), namely consensus investigation (via submodules msa and consensus), substrate mining (via submodule miner) and substrate biopathological pathway searching (via submodule pathfinder).

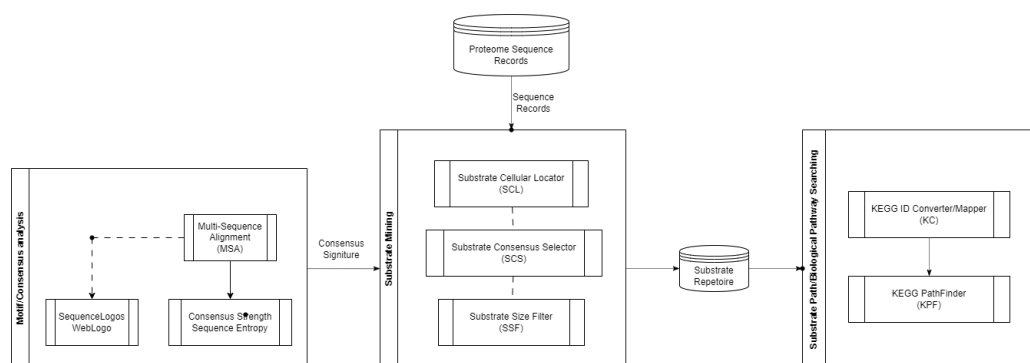


Figure 1: Substrateminor workflow

A typical workflow for adopting substrateminor to investigate substrate degradome for a given protease often involves the following three main stages:

Consensus investigation

Substrateminor provides a mechanism to visualise the cleavage consensus motif and calculate the strength of conservation at each site along the motif. The strength of conservation is calculated based on entropy, and substrateminor is also packaged with auxiliary tools to provide multi-sequence alignment (MSA) access to help prepare the input data for consensus investigation. Consensus visualisation is created by calling an implementation of weblogo (Crooks et al., 2004). (Figure 2)

As of the current release, substrateminor supports two popular multi-alignment tools for proteins, namely MUSCLE (default) (Edgar, 2004) and MAFFT (Katoh et al., 2002), while the Linux distribution supports additional implementation of Clustal0 (Sievers & Higgins, 2014). The cleavage motif consensus is calculated based on the frequency of amino acid residues at each position along the motif, and the visualisation is proportional to the information content at each site. The information content R at position i for protein is defined by Equation 1:

$$R_i = -\log_2(20) - H(X_i) \quad (1)$$

The strength of conservation is calculated based on Shannon entropy (Shannon, 1948) with the formulation in Equation 2, and the consensus motif is defined as the most frequent amino acid residue at each position along the motif.

$$H(X) = - \sum_{i=1}^n p_i \log_2 p_i \quad (2)$$

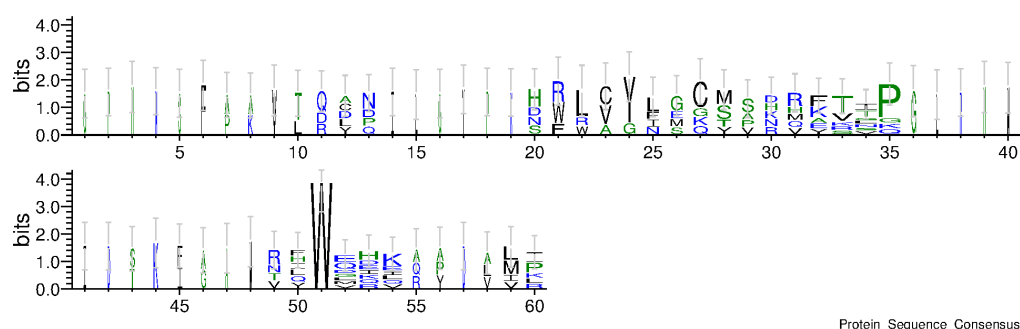


Figure 2: Consensus investigation - Visualisation Example

Substrate mining

Upon the identification of a cleavage consensus, substrateminer provides a mechanism to mine potential substrates based on three filter strategies:

1. **Cellular location:** filter potential substrates based on the cellular localisation of the intended target. For example, if the target protease is a secreted protease, the substrates mined should be extracellular proteins.
2. **Target Size:** filter potential substrates based on the size of the intended target. For example, if the target substrate is a small protease, the substrates mined should be small proteins or peptides of a specified size.
3. **Consensus:** filter potential substrates based on the conservation of the cleavage consensus. In the current release, both intra-protein cleavage (i.e., endopeptidase) and terminal cleavage (both C and N-terminal) (i.e., exopeptidase) consensus are supported.

Substrate pathobiological pathway searching

To further investigate the pathobiological relevance of novel substrates, substrateminer implements a search strategy to retrieve known biological processes and disease associations based on the *de facto* standard KEGG database (Kanehisa, 2019; Kanehisa & Goto, 2000). In the present release, the pathfinder module in the substrateminer provides three main functions, namely KEGG accession conversion, biological processes tracing and human disease mapping, where KEGG accession conversion enables a gateway to KEGG databases, where further information, including enzyme nomenclature and disease-related network databases, can be easily accessed. Biological processes tracing allows users to trace the biological processes associated with a given substrate, and human disease mapping enables users to map the potential substrates to known human diseases based on the KEGG disease database.

As of the current release, the pathfinder submodule queries all 580 KEGG molecular pathway maps and covers 1205 eukaryotes, 9375 bacteria and 449 archaea (Kanehisa & Goto, 2025). Nevertheless, the diseases covered are limited to the human context.

Limitations

The current release of substrateminor is limited to the investigation of protease substrates based on substrate primary sequence consensus in model organisms. In cases where proteases have very loose consensus, like metalloproteinase ADAM10 (Caescu et al., 2009) that is arguably best-known for its protective effects in prion diseases like mad cow disease in cattle, additional structural factors and contexts may need to be taken into consideration. Additionally, the efficacy of the substrateminor investigative power expanding beyond well-studied organisms is limited by the availability of a comprehensive proteome database for the underlying species. In the future, we aim to expand the scope of substrateminor to scrutinise substrate structural characteristics and provide a more comprehensive suite of tools for substrate investigation.

Conflict of Interests

The author declares no conflict of interest.

References

- Borges, P. H. O., Ferreira, S. B., & Silva, F. P. (2024). Recent advances on targeting proteases for antiviral development. *Viruses*, 16(3). <https://doi.org/10.3390/v16030366>
- Caescu, C. I., Jeschke, G. R., & Turk, B. E. (2009). Active-site determinants of substrate recognition by the metalloproteinases TACE and ADAM10. *Biochemical Journal*, 424(1), 79–88. <https://doi.org/10.1042/BJ20090549>
- Campbell, N. A., & Reece, J. B. (2005). *Biology*. Pearson, Benjamin Cummings. ISBN: 9780805371710
- Challen, B., & Cramer, R. (2022). Advances in ionisation techniques for mass spectrometry-based omics research. *PROTEOMICS*, 22(15-16), 2100394. <https://doi.org/10.1002/pmic.202100394>
- Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research*, 14(6), 1188–1190. <https://doi.org/10.1101/gr.849004>
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Han, X., Aslanian, A., & Yates, J. R. (2008). Mass spectrometry for proteomics. *Current Opinion in Chemical Biology*, 12(5), 483–490. <https://doi.org/10.1016/j.cbpa.2008.07.024>
- Kammerl, I. E., & Meiners, S. (2016). Proteasome function shapes innate and adaptive immune responses. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 311(2), L328–L336. <https://doi.org/10.1152/ajplung.00156.2016>
- Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11), 1947–1951. <https://doi.org/10.1002/pro.3715>
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Kanehisa, M., & Goto, S. (2025). *KEGG database current statistics*. Kanehisa Laboratories. <https://www.genome.jp/kegg/docs/statistics.html>
- Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059–3066. <https://doi.org/10.1093/nar/gkf436>

- Manasanch, E. E., & Orlowski, R. Z. (2017). Proteasome inhibitors in cancer therapy. *Nature Reviews Clinical Oncology*, 14, 417–433. <https://doi.org/10.1038/nrclinonc.2016.206>
- Papaneophytou, C. (2024). Breaking the chain: Protease inhibitors as game changers in respiratory viruses management. *International Journal of Molecular Sciences*, 25(15). <https://doi.org/10.3390/ijms25158105>
- Scott, L. J. (2017). Sitagliptin: A review in type 2 diabetes. *Drugs*, 77, 209–224. <https://doi.org/10.1007/s40265-016-0686-9>
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Sievers, F., & Higgins, D. G. (2014). Clustal omega. *Current Protocols in Bioinformatics*, 48(1), 3.13.1–3.13.16. <https://doi.org/10.1002/0471250953.bi0313s48>
- Sojka, D., Šnebergerová, P., & Robbertse, L. (2021). Protease inhibition—an established strategy to combat infectious diseases. *International Journal of Molecular Sciences*, 22(11). <https://doi.org/10.3390/ijms22115762>