

CREATION OF A NOVEL BIOLOGICAL DATABASE AND INFORMATION SYSTEM

FOR PIN2 PI: A DATA MINING APPROACH



सिद्धिमूलं प्रबन्धनम्
भा. प्र. सं. इन्दौर
IIM INDORE

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
EXECUTIVE FELLOW PROGRAMME IN MANAGEMENT

INDIAN INSTITUTE OF MANAGEMENT

INDORE

BY

Nikhilesh Kumar Yadav [2006, EFPM 03]

Date: September, 2021

Thesis Advisory Committee

Prof. Rajhans Mishra

[Chairman]

Prof. Shubhamoy Dey

[Member]

Prof. Mukul Gupta

[Member]

ABSTRACT

Technical advances in higher-throughput and lower-cost sequencing and genomics methods have led to an explosive growth of biological data, which, if not properly managed, will mean a waste of resources and efforts put in generating these data. Hence it is imperative to develop biological databases which would facilitate the collection, organization, analysis and sharing of biological data. An exponential increase in the acquisition of biological data and consequently its accumulation at larger scales poses unprecedented challenges in developing the biological databases. One such biological data resource is related to Plant Protease Inhibitors (PPIs). Protease Inhibitors (PIs) are proteins (sequence of amino acids) that have the potential to control proteolysis, an essential process in all living beings. Plant Protease Inhibitors (PPIs) are generally small proteins that offer a vital defense mechanism against insects and pests to the plants. There are various types of PIs present in plants, active against numerous proteases. Several families of PIs have been reported depending on specificity towards target proteases, their molecular mass and structure. Some of the PI families are Kunitz, Bowman-Birk, Squash, Potato type I inhibitors (PIN1) and potato type II (PIN2) protease inhibitors.

Numerous Plant PIs have been discovered, and their functions and families are known. However, new compounds are rapidly being discovered with uncharacterized functions and no knowledge about which family they belong to. In this research work using an amalgamation of existing Plant PIs database and in-house functional information about Plant PIs, we propose for preparing a novel specialized database of one such Plant PI known as **Potato Type Inhibitor-II family PIs** (Pin-II type PIs) plant protease inhibitors. Pin-II type PIs are well known as plant defense molecules against biotic stress. Also, these are potential molecules for the engineering of PIs because of features like simultaneous inhibition of multiple proteases, disulfide-bonded inhibitory domains and short reactive loop. However, this family of inhibitors has not been

explored due to limited annotated information in the available protein databases. Here, we have developed a database for Pin-II type PIs, consisting of manually collected and curated information about protein sequences of Pin-II type PIs. Precisely, the position of Inhibitory repeat domains, Linker regions, Reactive Center Loop, and disulfide linkages are mapped on the Pin-II PI sequence. This information related to Pin-II PIs is not mapped in any general or family-specific protein databases. In the current release of PINIR, we have annotated the protein sequences of 415 Pin-II type PI, downloaded from UniProtKB. We have identified and specified the number and position of 695 IRDs, 75 Linkers, 63 RCL and 10 disulfide bond patterns on the Pin-II PI sequences. We have also developed a web-based information system to facilitate searching, analyzing, and downloading information related to Pin-II type PIs. The database, together with the web-based information system, is called PINIR (**Pin-II type PIs Information Resource**). A comprehensive analysis of the PINIR database has been done using several existing statistical charts and custom-built interactive visualization tools for exploratory data analysis of this family. Since Pin- II type PIs show diversity in number and sequence of IRDs, using PINIR as a family-specific database of Pin-II type PIs will help explore this PI family and increase the understanding of its classification and functional diversification. PINIR would help identify the PIN2 PI's and predict their functions from the newly discovered compounds before the wet-lab experiments, which would eventually help in the conservation of capital and time. This database will be continuously updated with additional features and sequences of Pin-II type PIs to ensure that PINIR serves as a scientific resource for further research into Pin-II type PIs.

Database URL:<https://pinir.ncl.res.in/>

Keywords: Biological database, Protease Inhibitor, Pin-II, Bioinformatics, Knowledge discovery

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGEMENTS	iv
WORK APPRECIATION CERTIFICATE	v
DATA PERMISSION CERTIFICATE	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	xii
LIST OF TABLES	xiv
LIST OF ABBREVIATIONS	xv
1 CHAPTER 1: INTRODUCTION	1
1.1 Bioinformatics	1
1.2 Biological Databases	3
1.3 Present Investigation	5
1.3.1 Motivation.....	6
1.3.2 Aims	7
1.3.3 Scope.....	9
1.3.4 Impact	9
1.4 Guiding Framework	10
1.4.1 Features Elucidation.....	10
1.4.2 Data Collection	11
1.4.3 Information Extraction.....	12
1.5 Thesis Structure.....	13
2 CHAPTER 2: LITERATURE STUDY	15
2.1 Proteolytic Enzymes.....	15
2.2 Protease Inhibitors.....	16
2.3 Plant Protease Inhibitors (PPIs).....	16
2.3.1 Transgenic plants	17
2.4 PIs Classification.....	18

2.5	Potato type – II (Pin-II) PIs Family.....	19
2.6	Online Protein Databases	20
2.6.1	Pfam	20
2.6.2	Interpro.....	21
2.6.3	Merops	21
2.6.4	Plant PI.....	22
2.6.5	Uniprot KB.....	22
2.7	Summary	23
3	CHAPTER 3: DEVELOPMENT MODEL	24
3.1	Development Model.....	24
3.2	Data Sources.....	25
3.3	Data Preprocessing.....	28
3.3.1	Uniprot data processing	29
3.3.2	RCL processing.....	29
3.3.3	Linkers processing	30
3.3.4	Domains processing	30
3.3.5	Obtaining new Features	31
3.4	Data Integration.....	32
4	CHAPTER 4: RESULTS AND ANALYSIS	34
4.1	Database	34
4.2	Website.....	39
4.2.1	Browse	39
4.2.2	Data Analysis	41
4.2.3	Search and Download	41
4.3	Analysis.....	42
4.3.1	Occurrence and domain distribution of Pin-II type PI sequences.....	45
4.3.2	Linker diversity and distribution in Pin-II type PIs	49
4.3.3	RCL distribution and specificity	52
4.3.4	Disulfide bond variations in the Pin-II family	54
4.3.5	Features of Potato Inhibitor type II (Pin-II type) protein family	57
5	CHAPTER 5: DATA VISUALIZATION.....	60
5.1	Introduction	60
5.2	Visualization: Design Principles	62
5.2.1	Diagram First	62
5.2.2	Layout	62

5.2.3	Saliency	63
5.2.4	Negative Space.....	64
5.2.5	Gestalt Principles	65
5.2.6	Perceptual Accuracy	66
5.2.7	Small Multiples.....	68
5.3	Visualization – Design Elements	68
5.3.1	Data.....	68
5.3.2	Visual Encoding.....	69
5.3.3	Color	69
5.3.4	Interactivity	71
5.3.5	Headings and Labels	72
5.3.6	Axis and Scale.....	72
5.3.7	Charts	72
5.4	PINIR Visualization Techniques.....	73
5.4.1	Generic Visualizations	73
5.4.2	Custom Visualizations	86
6	CHAPTER 6: CONCLUSION	92
6.1	Key findings and Future recommendations	92
6.1.1	Pin-II PI Database	92
6.1.2	PINIR DB Data Analysis	93
6.1.3	Website and Visualization	94
6.2	Contributions.....	95
6.2.1	Theoretical Contribution.....	95
6.2.2	Practitioner’s Contributions	96
	APPENDICES	99
	Appendix A: Code	99
	A.1 OrganismParser.py	99
	A.2 RclResidueIdentifier.py	99
	A.3 RclFinder.py.....	99
	A.4 LinkerFinder.py	101
	A.5 DomainTypeIdentifier.py.....	102
	A.6 DomainFinder.py	103
	A.7 DSBondFinder.py	104
	A.8 SequenceAminoComposition.py	106
	Appendix B: PINIR Database.....	108

B.1 MST_PI_GeneralFeatures	108
B.2 MST_Taxonomy	108
B.3 DET_Sequence_Domains	108
B.4 MST_Domains	109
B.5 MST_DomainType	109
B.6 MST_Reactive_Loops.....	109
B.7 MST_Target_Protease.....	109
B.8 MST_Linkers	110
B.9 MST_LinkerType.....	110
B.10 DET_Cross_References	110
B.11 DET_Domain_Biochemical_Properties.....	110
B.12 DET_Domain_Biophysical_Properties.....	110
B.13 DET_Domain_DisulphideBonds	111
B.14 DET_DomainStructureFold	111
B.15 DET_Iso-ElectricPoints	111
B.16 DET_PI_GeneOntology	111
B.17 MST_GeneOntologyType.....	112
B.18 DET_SignalPeptide.....	112
B.19 DET_SpatioTemporal_Distribution.....	112
B.20 MST_Publication	112
B.21 DET_AminoComposition	112
Appendix C: PINIR Website	114
C.1 Home Page	114
C.2 Browse: Proteins (PI) Page	114
C.3 Proteins (PI) Details Page	115
C.4 Browse: Domains (IRD) Page.....	115
C.5 Browse: Reactive Loops (RCL) Page	116
C.6 Browse: Linkers Page	116
C.7 Browse: Taxonomy Page	117
C.8 Browse: Domain Architecture Page.....	117
C.9 Browse: Target Protease Page.....	118
C.10 Data Analysis: Proteins (PI) Page	118
C.11 Data Analysis: Domains (IRD) Page	119
C.12 Data Analysis: Reactive Loops (RCL) Page.....	119
C.13 Data Analysis: Linkers Page	120
C.14 Data Analysis: Target Protease Page	120
C.15 Search Page	121

C.16 Downloads Page.....	121
Appendix D: Case Study.....	122
REFERENCES.....	126

LIST OF FIGURES

Figure 1.1: Architecture Diagram	10
Figure 3.1: PINIR Development Model	25
Figure 3.2: Data Model of PINIR database	26
Figure 4.1(a): Database schema of Pin-II PI entity	35
Figure 4.1(b): Database schema of Domain entity	36
Figure 4.2: Main features of PINIR website.....	40
Figure 4.3: Database Statistics.....	44
Figure 4.4: Characteristics of Pin-II PIs.....	45
Figure 4.5: Distribution of Pin-II type PIs.....	46
Figure 4.6: Linker regions in Pin-II type PIs.	50
Figure 4.7: RCL distribution in Pin-II type PIs.....	54
Figure 4.8: Amino Acids distribution in Pin-II PIs.	55
Figure 5.1(a) The “Golden Section”	63
Figure 5.1(b)The “rule of thirds”	63
Figure 5.2 Examples of visual attributes to provide Saliency to design elements	64
Figure 5.3 Whitespace can be used to organize content and give the perception of hierarchy.....	64
Figure 5.4 Visual completion	65
Figure 5.5 Visual attribute ranking for encoding (a) Quantitative/ ordered data, and (b) Categorical data	66
Figure 5.6 Comparing by (I) Line Size and (II) Area Size	67

Figure 5.7(a) Use Hue for Categorical data.....	70
Figure 5.7(b) Use Lightness/ Saturation for quantitative data	70
Figure 5.7(c) Use divergent colormap for low/high data.....	70
Figure 5.8 Bar Chart: Genus wise distribution of sequences.....	75
Figure 5.9 Clustered Bar Chart: Top 8 Organism wise distribution of Target Protease	77
Figure 5.10 Pie Chart: Target Protease distribution in Domains	79
Figure 5.11 Box-and-Whiskers plot.....	81
Figure 5.12 Box Plot: Amino Acids percentage composition distribution.....	82
Figure 5.13 Heat Map: Top 8 Organisms wise distribution of RCL.....	84
Figure 5.14 PINIR Taxonomy Tree.....	88
Figure 5.15 PINIR Feature Viewer	91

LIST OF TABLES

Table 1.1: Available information about the Pin-II type PI sequences	7
Table 3.1: Data Sources for PINIR database	27
Table 4.1 Summary of the content of the tables in the PINIR database.....	38
Table 4.2: Occurrence of IRDs in PINIR database	47
Table 4.3: Species-wise distribution of IRDs (occurrence frequency >10).....	47
Table 4.4: Domain architectures in Pin-II type PI family.....	48
Table 4.5: Species distribution of multidomain Pin-II PIs.....	49
Table 4.6: Genus-wise distribution of linker regions.....	51
Table 4.7: Occurrence and distribution of RCL	53
Table 4.8: Distribution of IRDs according to dsBond type and correlation with linker types.....	56
Table 4.9: Genus-wise distribution of IRDs in dsBond types	56

LIST OF ABBREVIATIONS

PI	Protease Inhibitor
PPI	Plant Protease Inhibitors
PIN	Potato Inhibitor
Pin-II type PIs	Potato type Inhibitor-II family Protease Inhibitors
PIN2	Potato type 2 Inhibitors
PIN1	Potato type 1 Inhibitors
PINIR	Pin-II type PIs Information Resource
IRD	Inhibitory Repeat Domain
HMM	Hidden Markov Model
SCOP	Structural Classification of Proteins
PDB	Protein Data Bank
EMBL	European Molecular Biology Laboratory
RCL	Reactive Loops
GO	Gene Ontology
RDBMS	Relational Database Management System
AA	Amino Acid
PNG	Portable Network Graphics
SVG	Scalable Vector Graphics
CSV	Comma-separated values

REFERENCES

- Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, 27(1), 17–21.
<https://doi.org/10.1080/00031305.1973.10478966>
- Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506–D515. <https://doi.org/10.1093/nar/gky1049>
- Bertin, J. (1983). *Semiology of Graphics*. University of Wisconsin Press.
- Bioinformatics: Converting Data to Knowledge*. (2000). National Academies Press.
<https://doi.org/10.17226/9990>
- Bioinformatics* - *Wikipedia*. (n.d.). Retrieved July 23, 2020, from
<https://en.wikipedia.org/wiki/Bioinformatics>
- Birk, Y. (2003). *Plant protease inhibitors: significance in nutrition, plant protection, cancer prevention and genetic engineering*. Berlin: Springer-Verlag.
<https://books.google.com/books?hl=en&lr=&id=3TDxPqeOV34C&oi=fnd&pg=PA1&ots=dDA8hK3TAj&sig=JnJ8QipUYKPUNmN89aJGqbiYZ-U>
- Birney, E., & Clamp, M. (2004). Biological database design and implementation. *Briefings in Bioinformatics*, 5(1), 31–38. <https://doi.org/10.1093/bib/5.1.31>
- Bostock, Michael, Ogievetsky, V., & Heer, J. (2011). D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2301–2309.
<https://doi.org/10.1109/TVCG.2011.185>
- Bostock, Mike. (n.d.). *Collapsible Tree / D3 / Observable*. Retrieved January 2, 2021, from
<https://observablehq.com/@d3/collapsible-tree>
- Bry, F., & Kröger, P. (2003). A Computational Biology database digest: Data, data analysis,

- and data management. *Distributed and Parallel Databases*, 13(1), 7–42.
<https://doi.org/10.1023/A:1021540705916>
- Can, T. (2014). Introduction to bioinformatics. *Methods in Molecular Biology*, 1107, 51–71.
https://doi.org/10.1007/978-1-62703-748-8_4
- Ceroni, A., Passerini, A., Vullo, A., & Frasconi, P. (2006). DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Research*, 34(suppl_2), W177–W181. <https://doi.org/10.1093/nar/gkl266>
- Clemente, M., Corigliano, M. G., Pariani, S. A., Sánchez-López, E. F., Sander, V. A., & Ramos-Duarte, V. A. (2019). Plant serine protease inhibitors: Biotechnology application in agriculture and molecular farming. In *International Journal of Molecular Sciences* (Vol. 20, Issue 6, p. 1345). MDPI AG. <https://doi.org/10.3390/ijms20061345>
- Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716), 828–833.
<https://doi.org/10.1126/science.229.4716.828>
- ColorBrewer: Color Advice for Maps*. (n.d.). Retrieved January 3, 2021, from <https://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3>
- Consortium, T. U. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506–D515. <https://doi.org/10.1093/nar/gky1049>
- Council, N. R. (2000). *Bioinformatics: Converting Data to Knowledge* (R. Pool & J. Esnayra (Eds.)). The National Academies Press. <https://doi.org/10.17226/9990>
- D3.js - Data-Driven Documents*. (n.d.). Retrieved January 2, 2021, from <https://d3js.org/>
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L.,

- Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1), D427–D432. <https://doi.org/10.1093/nar/gky995>
- Fan, S. G., & Wu, G. J. (2005). Characteristics of plant proteinase inhibitors and their applications in combating phytophagous insects. In *Botanical Bulletin of Academia Sinica* (Vol. 46, Issue 4, pp. 273–292). <https://doi.org/10.7016/BBAS.200510.0273>
- Farady, C. J., & Craik, C. S. (2010). Mechanisms of Macromolecular Protease Inhibitors. In *ChemBioChem* (Vol. 11, Issue 17, pp. 2341–2346). <https://doi.org/10.1002/cbic.201000442>
- Few, S. (2007). *Data Visualization - Past, Present, and Future*.
- Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., ... Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research*, 45(D1), D190–D199. <https://doi.org/10.1093/nar/gkw1107>
- Gartia, J., Anangi, R., Joshi, R. S., Giri, A. P., King, G. F., Barnwal, R. P., & Chary, K. V. R. (2020). NMR structure and dynamics of inhibitory repeat domain variant 12, a plant protease inhibitor from *Capsicum annuum*, and its structural relationship to other plant protease inhibitors. *Journal of Biomolecular Structure and Dynamics*, 38(5), 1388–1397. <https://doi.org/10.1080/07391102.2019.1607559>
- Gehlenborg, N., & Wong, B. (2012a). Points of view: Heat maps. In *Nature Methods* (Vol. 9, Issue 3, p. 213). Nat Methods. <https://doi.org/10.1038/nmeth.1902>
- Gehlenborg, N., & Wong, B. (2012b). Points of view: Mapping quantitative data to color. In

Nature Methods (Vol. 9, Issue 8, p. 769). *Nat Methods*.
<https://doi.org/10.1038/nmeth.2134>

github. (2021). *GitHub*. <https://github.com/>

Habib, H., & Fazili, K. M. (2006). Biotechnology and molecular biology reviews. In *Biotechnology and Molecular Biology Reviews* (Vol. 2, Issue 3). Academic Journals.
<https://academicjournals.org/journal/BMBR/article-abstract/8EB195410993>

Habib, H., & Fazili, K. M. (2007). Plant protease inhibitors : a defense strategy in plants. *Biotechnology and Molecular Biology Review*, 2(3), 68–85. <https://doi.org/ISSN 1538-2273>

Helmy, M., Crits-Christoph, A., & Bader, G. D. (2016). Ten Simple Rules for Developing Public Biological Databases. In *PLoS Computational Biology* (Vol. 12, Issue 11, p. e1005128). Public Library of Science. <https://doi.org/10.1371/journal.pcbi.1005128>

HMMER. (n.d.). Retrieved May 24, 2020, from <http://hmmer.org/>

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., ... Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Research*, 37(Database issue), D211-5. <https://doi.org/10.1093/nar/gkn785>

InterPro. (n.d.). Retrieved March 6, 2020, from <https://www.ebi.ac.uk/interpro/>

Jamal, F., Pandey, P. K., Singh, D., & Khan, M. Y. (2013). Serine protease inhibitors in plants: Nature's arsenal crafted for insect predators. In *Phytochemistry Reviews* (Vol. 12, Issue 1, pp. 1–34). Springer Netherlands. <https://doi.org/10.1007/s11101-012-9231-y>

Kirk, A. (2016). *Data Visualisation: A Handbook for Data Driven Design*. Sage Publications

Ltd.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90).

Koiwa, H., Bressan, R. A., & Hasegawa, P. M. (1997). Regulation of protease inhibitors and plant defense. *Trends in Plant Science*, 2(10), 379–384. [https://doi.org/10.1016/s1360-1385\(97\)90052-2](https://doi.org/10.1016/s1360-1385(97)90052-2)

Kong, L., & Ranganathan, S. (2008). Tandem duplication, circular permutation, molecular adaptation: how Solanaceae resist pests via inhibitors. *BMC Bioinformatics*, 9 Suppl 1, S22. <https://doi.org/10.1186/1471-2105-9-S1-S22>

Kozłowski, L. P. (2016). IPC – Isoelectric Point Calculator. *Biology Direct*, 11(1), 55. <https://doi.org/10.1186/s13062-016-0159-9>

Kroger, P. (2001). Molecular Biology Data : Database Overview , Modelling Issues , and Perspectives. *Database*. <https://www.pms.ifi.lmu.de/publikationen/diplomarbeiten/Peer.Kroeger/diplomarbeit.pdf>

Krzywinski, M. (2013a). Points of view: Axes, ticks and grids. In *Nature Methods* (Vol. 10, Issue 3, p. 183). <https://doi.org/10.1038/nmeth.2337>

Krzywinski, M. (2013b). Points of view: Labels and callouts. In *Nature Methods* (Vol. 10, Issue 4, p. 275). *Nat Methods*. <https://doi.org/10.1038/nmeth.2405>

Laskowski, M., & Kato, I. (1980). Protein Inhibitors of Proteinases. *Annual Review of*

Biochemistry, 49(1), 593–626. <https://doi.org/10.1146/annurev.bi.49.070180.003113>

Lawrence, P. K., & Koundal, K. R. (2002). Plant protease inhibitors in control of phytophagous insects. In *Electronic Journal of Biotechnology* (Vol. 5, Issue 1, pp. 93–109). <https://doi.org/10.2225/vol5-issue1-fulltext-3>

Leo, F. D. (2002). PLANT-PIs: a database for plant protease inhibitors and their genes. *Nucleic Acids Research*, 30(1), 347–348. <https://doi.org/10.1093/nar/30.1.347>

Leo, F. D., Volpicella, M., Licciulli, F., Liuni, S., Gallerani, R., & Ceci, L. R. (2002). PLANT-PIs: a database for plant protease inhibitors and their genes. *Nucleic Acids Research*, 30(1), 347–348. <https://doi.org/10.1093/nar/30.1.347>

Luo, M., Ding, L. W., Ge, Z. J., Wang, Z. Y., Hu, B. L., Yang, X. B., Sun, Q. Y., & Xu, Z. F. (2012). The characterization of SaPIN2b, a plant trichome-localized proteinase inhibitor from *Solanum americanum*. *International Journal of Molecular Sciences*, 13(11), 15162–15176. <https://doi.org/10.3390/ijms131115162>

Mackinlay, J., & Mackinlay, J. (1986). Automating the Design of Graphical Presentations of Relational Information. *ACM TRANSACTIONS ON GRAPHICS*, 5, 110–141. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.95.1731>

MEROPS - the Peptidase Database. (n.d.). Retrieved March 6, 2020, from <https://www.ebi.ac.uk/merops/inhibitors/>

Midway, S. R. (2020). Principles of Effective Data Visualization. In *Patterns* (Vol. 1, Issue 9, p. 100141). Cell Press. <https://doi.org/10.1016/j.patter.2020.100141>

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., ... Finn, R. D. (2018).

- InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1), D351–D360. <https://doi.org/10.1093/nar/gky1100>
- Munzner, T. (2014). Visualization Analysis and Design. In *Visualization Analysis and Design*. A K Peters/CRC Press. <https://doi.org/10.1201/b17511>
- O’Donoghue, S. I., Baldi, B. F., Clark, S. J., Darling, A. E., Hogan, J. M., Kaur, S., Maier-Hein, L., McCarthy, D. J., Moore, W. J., Stenau, E., Swedlow, J. R., Vuong, J., & Procter, J. B. (2018). Visualization of Biomedical Data. *Annual Review of Biomedical Data Science*, 1(1), 275–304. <https://doi.org/10.1146/annurev-biodatasci-080917-013424>
- O’Donoghue, S. I., Gavin, A. C., Gehlenborg, N., Goodsell, D. S., Hériché, J. K., Nielsen, C. B., North, C., Olson, A. J., Procter, J. B., Shattuck, D. W., Walter, T., & Wong, B. (2010). Visualizing biological data—now and in the future. In *Nature Methods* (Vol. 7, Issue 3, pp. S2–S4). Nat Methods. <https://doi.org/10.1038/nmeth.f.301>
- Oliva, M. L. V., & Sampaio, M. U. (2009). Action of plant proteinase inhibitors on enzymes of physiopathological importance. *Anais Da Academia Brasileira de Ciencias*, 81(3), 615–621. <https://doi.org/10.1590/S0001-37652009000300023>
- Paiva, P. M. G., Pontual, E. V, Coelho, L. C. B. B., & Napoleão, T. H. (2013). Protease inhibitors from plants: Biotechnological insights with emphasis on their effects on microbial pathogens. *Microbial Pathogens and Strategies for Combating Them: Science, Technology and Education, Figure 1*, 641–649.
- Paladin, L., Schaeffer, M., Gaudet, P., Zahn-Zabal, M., Michel, P. A., Piovesan, D., Tosatto, S. C. E., & Bairoch, A. (2020). The Feature-Viewer: A visualization tool for positional annotations on a sequence. *Bioinformatics*, 36(10), 3244–3245. <https://doi.org/10.1093/bioinformatics/btaa055>

- Rawlings, N. D. (2010). Peptidase inhibitors in the MEROPS database. In *Biochimie* (Vol. 92, Issue 11, pp. 1463–1483). Elsevier. <https://doi.org/10.1016/j.biochi.2010.04.013>
- Rawlings, N. D., Barrett, A. J., & Finn, R. (2016). Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, *44*(D1), D343-50. <https://doi.org/10.1093/nar/gkv1118>
- Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research*, *46*(D1), D624–D632. <https://doi.org/10.1093/nar/gkx1134>
- Rawlings, N. D., Tolle, D. P., & Barrett, A. J. (2004). Evolutionary families of peptidase inhibitors. In *Biochemical Journal* (Vol. 378, Issue 3, pp. 705–716). Portland Press. <https://doi.org/10.1042/BJ20031825>
- Rawlings, N. D., Waller, M., Barrett, A. J., & Bateman, A. (2014). MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, *42*(D1). <https://doi.org/10.1093/nar/gkt953>
- Rustgi, S., Boex-Fontvieille, E., Reinbothe, C., von Wettstein, D., & Reinbothe, S. (2018). The complex world of plant protease inhibitors: Insights into a Kunitz-type cysteine protease inhibitor of *Arabidopsis thaliana*. In *Communicative and Integrative Biology* (Vol. 11, Issue 1, p. e1368599). Taylor & Francis. <https://doi.org/10.1080/19420889.2017.1368599>
- Saikhedkar, N. S., Joshi, R. S., Bhoite, A. S., Mohandasan, R., Yadav, A. K., Fernandes, M., Kulkarni, K. A., & Giri, A. P. (2018). Tripeptides derived from reactive centre loop of potato type II protease inhibitors preferentially inhibit midgut proteases of *Helicoverpa armigera*. *Insect Biochemistry and Molecular Biology*, *95*, 17–25. <https://doi.org/10.1016/J.IBMB.2018.02.001>

- Schirra, H. J., Guarino, R. F., Anderson, M. A., & Craik, D. J. (2010). Selective Removal of Individual Disulfide Bonds within a Potato Type II Serine Proteinase Inhibitor from *Nicotiana alata* Reveals Differential Stabilization of the Reactive-Site Loop. *Journal of Molecular Biology*, 395(3), 609–626. <https://doi.org/10.1016/j.jmb.2009.11.031>
- Tamhane, V. A., Mishra, M., Mahajan, N. S., Gupta, V. S., & Giri, A. P. (2012). Plant Pin-II family proteinase inhibitors: structural and functional diversity. *Funct Plant Sci Biotechnol*, 6, 42–58. <http://merops.sanger.ac.uk>
- Thomas Triplet, & Gregory Butler. (2011). Systems biology warehousing: challenges and strategies toward effective data integration. *Proc. 3rd International Conference on Advances in Databases, Knowledge, and Data Applications, St. Maarten. IARIA*, 34–40.
- Unwin, A. (2020). Why is Data Visualization Important? What is Important in Data Visualization? *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.8ae4d525>
- Wes McKinney. (2012). *Python for Data Analysis [Book]*. O'Reilly Media, Inc.
- What is bioinformatics? | Bioinformatics for the terrified.* (n.d.). Retrieved July 23, 2020, from <https://www.ebi.ac.uk/training-beta/online/courses/bioinformatics-terrified/what-bioinformatics/>
- Wong, B. (2010a). Design of data figures. *Nature Methods*, 7(9), 665. <https://doi.org/10.1038/nmeth0910-665>
- Wong, B. (2010b). Points of view: Color coding. In *Nature Methods* (Vol. 7, Issue 8, p. 573). Nature Publishing Group. <https://doi.org/10.1038/nmeth0810-573>
- Wong, B. (2010c). Saliency. *Nature Methods*, 7(10), 773. <https://doi.org/10.1038/nmeth1010-773>

- Wong, B. (2010d). Points of View: Gestalt principles (Part 1). In *Nature Methods* (Vol. 7, Issue 11, p. 863). Nat Methods. <https://doi.org/10.1038/nmeth1110-863>
- Wong, B. (2010e). Points of View: Gestalt principles (Part 2). In *Nature Methods* (Vol. 7, Issue 12, p. 941). <https://doi.org/10.1038/nmeth1210-941>
- Wong, B. (2011a). Negative space. *Nature Methods*, 8(1), 5. <https://doi.org/10.1038/nmeth0111-5>
- Wong, B. (2011b). Points of view: Typography. *Nature Methods*, 8(4), 277. <https://doi.org/10.1038/nmeth0411-277>
- Wong, B. (2011c). Salience to relevance. *Nature Methods*, 8(11), 889. <https://doi.org/10.1038/nmeth.1762>
- Wong, B. (2011d). Points of view: Layout. In *Nature Methods* (Vol. 8, Issue 10, p. 783). Nat Methods. <https://doi.org/10.1038/nmeth.1711>
- Wong, B. (2012). Visualizing biological data. *Nature Methods*, 9(12), 1131. <https://doi.org/10.1038/nmeth.2258>
- Wong, B., & Kjægaard, R. S. (2012). Pencil and paper. *Nature Methods*, 9(11), 1037. <https://doi.org/10.1038/nmeth.2223>
- Yadav, N. K., Saikhedkar, N. S., & Giri, A. P. (2021). PINIR: a comprehensive information resource for Pin-II type protease inhibitors. *BMC Plant Biology*, 21(1), 267. <https://doi.org/10.1186/s12870-021-03027-0>
- zenodo. (2021, May 8). ZENODO. <https://zenodo.org/>.
- Zhang, S. Y., & Liu, S. L. (2013). Bioinformatics. In *Brenner's Encyclopedia of Genetics: Second Edition* (pp. 338–340). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-374984->

0.00155-8

Zou, D., Ma, L., Yu, J., & Zhang, Z. (2015). Biological databases for human research. In *Genomics, Proteomics and Bioinformatics* (Vol. 13, Issue 1, pp. 55–63). Elsevier.
<https://doi.org/10.1016/j.gpb.2015.01.006>