MolEvolvR: A web-app for characterizing proteins using molecular evolution and phylogeny

Faisal S Alquaddoomi[1,#], Joseph T Burke[2,3,#], Lo Sosinski[3,5,#], David A Mayer[1], Evan P Brenner[1], Samuel Z Chen[3,4], Jacob D Krol[1], Ethan P Wolfe[1,5], Vince P Rubinetti[1], Shaddai Amolitos[6], Kellen M Reason[3], John B Johnston[7], Janani Ravi[1,3*]

[1]Department of Biomedical Informatics, Center for Health Artificial Intelligence, University of Colorado Anschutz, Aurora, CO 80045; [2]Genomics and Molecular Genetics Undergraduate Program, Michigan State University, East Lansing, MI 48824; [3]Department of Pathobiology and Diagnostic Investigation, Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824; [4]Computer Science Undergraduate Program, Michigan State University, East Lansing, MI 48824; [5]Biochemistry and Biotechnology Undergraduate Program, Michigan State University, East Lansing, MI 48824; [6]School of Medicine, University of Colorado Anschutz, Aurora, CO 80045; [7]Center for Genomics-Enabled Plant Sciences, Michigan State University, East Lansing, MI 48824.
[#]Co-primary authors contributed equally. *Corresponding author: janani.ravi@cuanschutz.edu.

Abstract

Studying proteins through the lens of evolution can reveal features such as conserved domains, lineage-specific variants, and co-occurring domain architectures in phylogenetic context across all superkingdoms. *MolEvolvR* enables researchers to conduct such evolution-focused studies to generate testable hypotheses about protein function and evolution. *MolEvolvR* is a novel web-app allowing researchers to visualize the molecular evolution of their proteins of interest in a phylogenetic context across the tree of life. It accepts multiple input formats – protein/domain sequences, homologous proteins, or domain scans – and, using a general-purpose computational workflow, returns detailed homolog data and dynamic graphical summaries (e.g., phylogenetic trees, multiple sequence alignments, domain architectures, domain proximity networks, phyletic spreads, co-occurrence patterns across lineages). *MolEvolvR* performs domain-centric searches to capture remote homologs that are missed by full-length searches, integrates domain architecture evolution with phyletic distribution analyses, and provides evolutionary context visualizations that reveal lineage-specific adaptations versus those that are broadly conserved. Thus, *MolEvolvR* is a powerful, easy-to-use web interface for computational protein characterization. The web-app can be accessed here: https://jravilab.org/molevolvr.

Keywords

Molecular evolution; Phylogenetics; Domain architectures; Homology; Protein characterization; Comparative genomics; Web application

The rate at which protein families are being discovered far outpaces that at which molecular or biochemical functions are assigned to proteins[1]. This gap impedes scientific understanding of critical cellular processes such as molecular pathogenesis and stress response. Identifying and characterizing the complete molecular systems involved in such processes requires functional knowledge of the proteins involved. Numerous studies[2–18] have demonstrated the power of phylogenetic and molecular evolutionary analysis in determining the molecular functions of such proteins. A variety of individual tools[19–45] exist for protein characterization, including sequence similarity searches or ortholog detection, delineating co-occurring domains (domain architectures), and building multiple sequence alignments and phylogenetic trees. However, these tools operate in isolation, requiring users to manually integrate and interpret results within an evolutionary context. Despite providing complementary insight, these tools are disjointed and often require both technical skill and background knowledge for meaningful use. Biologists need a means to functionally characterize their unknown/understudied protein(s) of interest by exhaustively identifying relevant homologs with shared motifs, domains, and domain architectures. Moreover, homology searches using full-length protein sequences often fail to identify remote homologs that share only individual functional domains.

Here, we present *MolEvolvR*, a unified, user-friendly, and accessible web-based framework that addresses these limitations by utilizing data from all taxonomic domains to analyze biological sequences of interest in depth (**Fig. 1**; accessible at jravilab.org/molevolvr). Given one or more proteins of interest from the user, *MolEvolvR* performs both protein- and domain-centric searches to capture remote homologs, reconstructs domain architecture evolution across the tree of life, and provides evolutionary context visualizations that distinguish broadly conserved features from lineage-specific adaptations. These findings, spanning multiple taxonomic scales, are effectively summarized and visualized to generate functional hypotheses. The computational evolutionary framework underlying *MolEvolvR* is written using custom R[46–61] and shell scripts. The web application is built with an R/Shiny[48,62,63] framework with additional user interface customizations in HTML, JavaScript, and CSS, and has been tested on Chrome, Firefox, and Safari browsers on Mac, Windows, and Linux operating systems.

Consider a researcher starting with a protein of unknown function, either identified through a genetic screen for a phenotype or an uncharacterized protein in a genomic segment of interest [**Fig. S1**]. In Step 1, *MolEvolvR* resolves this query protein into its constituent domains and uses the protein and each of its domains for iterative homology searches[20,21] across the tree of life[64–67] [**Fig. S2**]. This divide-and-conquer strategy is domain-sensitive by design and captures remote homologs missed by homology searches using full-length protein sequences alone. In Step 2, *MolEvolvR* characterizes the protein and each of its homologs by reconstructing their domain architectures and delineating putative molecular function [**Fig. S3**], combining: 1) sequence alignment and clustering algorithms for domain detection[19,24,25,68]; 2) profile matching against protein domain and orthology databases[32,33,69–72]; and 3) prediction algorithms for signal peptides[33,73], transmembrane regions[33,74–76], cellular localization[33,75], and secondary/tertiary structures[33,77–80]. In Step 3, *MolEvolvR* maps each of these homologs and their domain architectures to their lineages to further tease out broadly conserved vs. phylum-specific

signatures, placing these findings in the context of evolution [**Fig. S4**]. This analysis illustrates the breadth of *MolEvolvR*'s analytic capabilities: the results entail a detailed molecular characterization of the initial query protein(s) and their homologs, integrated across the superkingdoms of life, revealing both universally conserved features and any lineage-specific functional variants.

Beyond single protein inputs, *MolEvolvR* is versatile, accommodating a variety of inputs for investigating wide-ranging questions and producing diverse output summaries and visualizations in easy-to-access formats. Queries can include protein/domain sequences of single- or multi-protein operons (FASTA, NCBI/UniProt protein accession numbers), homologous proteins (pre-computed MSA or web/command-line BLAST output), or motif/domain scans (InterProScan output) [**Fig. 1B**]. *MolEvolvR* tailors analyses to answer a variety of questions, e.g., determining protein or functional domain features restricted to certain pathogenic groups using the organism filter to discover virulence factors [**Fig. 1C**; **Fig. S1**]. Finally, *MolEvolvR* generates output types spanning a complete set of homologs or phylogenetic trees, the domain architectures of a query protein, or the most common partner domains [previews in **Fig. 1A**]. Along with tables and visualizations, *MolEvolvR* also provides a dashboard with summary statistics and graphical summaries integrating these results in the context of evolution [**Fig. S1**]: i) structure-based multiple sequence alignments and phylogenetic trees; ii) domain proximity networks from all co-occurring domains (across homolog domain architectures) consolidating within and across query proteins; iii) phyletic spreads of homologs and their domain architectures; and iv) co-occurrence patterns and relative occurrences of domain architectures across lineages [**Fig. 1A**]. The domain-centric searches trace the evolution of the proteins/domains of interest, even across remote homologs. The web-app contains detailed documentation about all these options, including case studies (see *Supplementary Material*) and frequently asked questions (FAQs). In addition to easy access to a local, in-browser history of user-submitted jobs, *MolEvolvR* can, optionally, send users a detailed description and status update of their submitted jobs via email. Finally, all figures, tables, and reports can be downloaded by the user.

MolEvolvR's integrative approach has enabled the discovery of functional properties across diverse protein systems. Analysis of the phage shock protein (PSP) stress response system identified ~20,000 homologs and ~200 novel domain architectures, revealing evolutionary variants adapted to different membrane stress challenges across bacteria, archaea, and eukaryotes[2]. Domain-centric searches captured homologs sharing core functional domains (e.g., PspA) despite <30% overall sequence identity, while phyletic spread analyses distinguished universally conserved components from lineage-specific adaptations. A specific instance and adaptation of *MolEvolvR*, applied to delineate the PSP stress response system in great detail across the tree of life, is available at https://jravilab.org/psp[2] [see *Supplementary Case Study*]. *MolEvolvR* is a generalized web server of this web-app. To demonstrate its broad applicability, we have successfully applied the approach underlying *MolEvolvR* to discover and functionally characterize several systems, including proteins/operons in zoonotic pathogens, e.g., two nutrient acquisition systems in *Staphylococcus aureus*[4,5], a novel phage defense

system in *Vibrio cholerae* (informing functional validation experiments)[6] surface layer proteins in *Bacillus anthracis*[7], helicase operators in bacteria[8], CARD[81] antimicrobial resistance genes and uncharacterized proteins associated with AMR by our machine learning models[82], and internalins in *Listeria* spp[9]. We have included a few pre-loaded examples in *MolEvolvR* for users to explore, in addition to help pages and FAQ. Researchers can submit requests for new analysis or visualization features through *MolEvolvR*.

Thus, *MolEvolvR* (jravilab.org/molevolvr) is a flexible, user-friendly, and powerful interactive web tool that enables researchers of any level of bioinformatics experience to bring molecular evolution and phylogeny to bear on their proteins of interest and provide evolutionary context and domain organization that inform functional hypotheses and guide experimental validation.

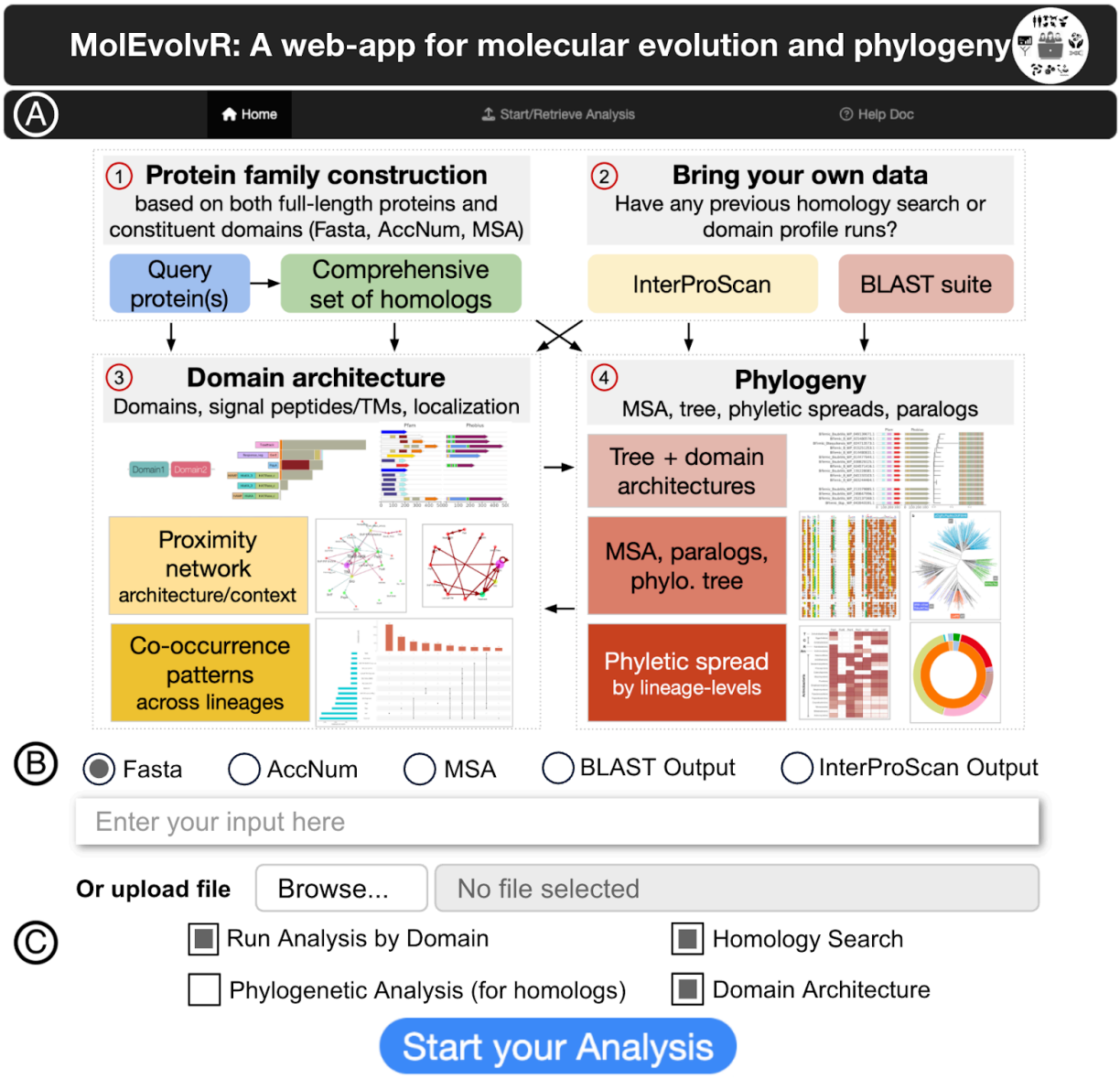

Figure 1. Overview of *MolEvolvR*. A. *MolEvolvR* allows users to start with protein(s) of interest and perform the complete analysis (1+3+4), only protein characterization (1+3), or only homology searches (1+4); or start with external outputs from BLAST or InterProScan for further analysis, summarization, and visualization (2+3+4). *MolEvolvR* is interactive, queryable, and customizable. **B.** Multiple input options in *MolEvolvR*: FASTA, NCBI/UniProt accession numbers, pre-computed multiple sequence alignment, MSA (in FASTA formats), analysis outputs from an external web or command-line BLAST or InterProScan runs. **C.** The different analysis options available in *MolEvolvR* depend on the chosen input formats in **1A**, **1B**. See *Supplementary case study* and web-app help docs (https://jravilab.org/molevolvr/?r=&p=help) for further details on versatile input/output options and data summarizations and visualizations.

Acknowledgments

We would like to thank members of the JRaviLab for testing the web-app at various stages and providing the authors with several iterations of constructive feedback. We are also profoundly grateful to Krishnan Raghunathan, Arjun Krishnan, Emily Meyer, Jerome McKay, Premal Shah, Dave Bunten, and members of the Krishnan lab (Kayla Johnson and Nathaniel Hawkins) for their early feedback on *MolEvolvR* and detailed comments on the manuscript. We have benefited from several diverse use cases and challenges brought to us by our collaborators that helped fine-tune the functionality of the web-app (inline citations), and the larger *MolEvolvR* user community that has engaged with us to keep the app current, relevant, and feature-rich.

Funding

We would like to thank our funding sources: University of Colorado Anschutz start-up funds, Endowed Research Funds from the College of Veterinary Medicine, Michigan State University, NSF-funded BEACON funding support awarded to JR; NSF-funded REU-ACRES summer scholarship to SZC; NIH NIAID U01AI176414 to JR; NIH NLM T15LM009451 to EPB.

Author Contributions

JR conceived and designed the study; JTB, LS, SZC, JDK, EPB, and JR acquired the data and performed all the analyses; FSA, JTB, LS, EPB, DAM, JDK, JR, FSA, and DAM made the figures and tables; FSA, JTB, LS, DAM, SZC, JDK, JR wrote the backend code; FSA, JTB, DAM, SZC, JDK, VPR, JR built the R/Shiny web-app; FSA, DAM, JBJ, KMR, and SA set up the server that hosts the web-app; FSA, EPB, DAM, EPW, JTB, VPR, JDK, LS, JR tested the web-app rigorously; EPB, EPW, FSA, DAM, JTB, JDK, LS, JR worked on the help documentation; LS and JR wrote the first draft of the manuscript; EPB, JTB, JDK, LS, EPW, and JR revised the manuscript.

Data Availability and Reuse

All the data, analyses, and visualizations are available in our interactive and queryable web application: http://jravilab.org/molevolvr. The web-app is licensed under the MIT license.

References

1. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, e1000605 (2009).
2. Ravi, J. *et al.* The phage shock protein (PSP) envelope stress response: discovery of novel partners and evolutionary history. *mSystems* **9**, e00847-23 (2024).
3. Ravi, J., Anantharaman, V., Aravind, L. & Gennaro, M. L. Variations on a theme: evolution of the phage-shock-protein system in Actinobacteria. *Antonie Van Leeuwenhoek* **111**, 753–760 (2018).
4. Lensmire, J. M. *et al.* The Staphylococcus aureus Cystine Transporters TcyABC and TcyP Facilitate Nutrient Sulfur Acquisition during Infection. *Infect. Immun.* **88**, (2020).
5. Lensmire, J. M. *et al.* The glutathione import system satisfies the Staphylococcus aureus nutrient sulfur requirement and promotes interspecies competition. *PLOS Genetics* **19**, e1010834 (2023).
6. Hsueh, B. Y. *et al.* Phage defence by deaminase-mediated depletion of deoxynucleotides in bacteria. *Nat Microbiol* https://doi.org/10.1038/s41564-022-01162-4 (2022) doi:10.1038/s41564-022-01162-4.
7. Ravi, J. & Fioravanti, A. S-layers: The Proteinaceous Multifunctional Armors of Gram-Positive Pathogens. *Front Microbiol* **12**, 663468 (2021).
8. Blaine, H. C., Burke, J. T., Ravi, J. & Stallings, C. L. DciA Helicase Operators Exhibit Diversity across Bacterial Phyla. *J Bacteriol* **204**, e0016322 (2022).
9. Conner, K. N., Burke, J. T., Ravi, J. & Hardy, J. W. Novel internalin P homologs in Listeria. *Microb Genom* **8**, mgen000828 (2022).
10. Manganelli, R. & Gennaro, M. L. Protecting from Envelope Stress: Variations on the Phage-Shock-Protein Theme. *Trends Microbiol.* **25**, 205–216 (2017).
11. Datta, P. *et al.* The Psp system of Mycobacterium tuberculosis integrates envelope stress-sensing and envelope-preserving functions. *Mol. Microbiol.* **97**, 408–422 (2015).
12. Kaur, G., Burroughs, A. M., Iyer, L. M. & Aravind, L. Highly regulated, diversifying NTP-dependent biological conflict systems with implications for the emergence of multicellularity. *Elife* **9**, (2020).
13. Shmakov, S. A., Makarova, K. S., Wolf, Y. I., Severinov, K. V. & Koonin, E. V. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E5307–E5316 (2018).
14. Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V. & Aravind, L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol. Direct* **8**, 15 (2013).
15. Koonin, E. V. Systemic determinants of gene evolution and function. *Mol. Syst. Biol.* **1**, 2005.0021 (2005).
16. Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136 (2003).
17. Vicedomini, R., Blachon, C., Oteri, F. & Carbone, A. MyCLADE: a multi-source domain annotation server for sequence functional exploration. *Nucleic Acids Research* **49**, W452–W458 (2021).
18. Lee, B. & Lee, D. DAhunter: a web-based server that identifies homologous proteins by comparing domain architecture. *Nucleic Acids Res.* **36**, W60-64 (2008).
19. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
20. Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7**, 12 (2012).
21. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
22. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

23. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
24. Magis, C. *et al.* T-Coffee: Tree-based consistency objective function for alignment evaluation. *Methods Mol. Biol.* **1079**, 117–129 (2014).
25. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
26. Lassmann, T. Kalign 3: multiple sequence alignment of large data sets. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz795 (2019) doi:10.1093/bioinformatics/btz795.
27. Gumerov, V. M. & Zhulin, I. B. TREND: a platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkaa243 (2020) doi:10.1093/nar/gkaa243.
28. Klimchuk, O. I. *et al.* COGNAT: a web server for comparative analysis of genomic neighborhoods. *Biol. Direct* **12**, 26 (2017).
29. Shmakov, S. A. *et al.* Systematic prediction of functionally linked genes in bacterial and archaeal genomes. *Nat Protoc* **14**, 3013–3031 (2019).
30. Adebali, O. & Zhulin, I. B. Aquerium: A web application for comparative exploration of domain-based protein occurrences on the taxonomically clustered genome tree. *Proteins* **85**, 72–77 (2017).
31. Persson, E., Kaduk, M., Forslund, S. K. & Sonnhammer, E. L. L. Domainoid: domain-oriented orthology inference. *BMC Bioinformatics* **20**, 523 (2019).
32. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344–D354 (2021).
33. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
34. Altenhoff, A. M. *et al.* OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res* **49**, D373–D379 (2021).
35. Cosentino, S. & Iwasaki, W. SonicParanoid: fast, accurate and easy orthology inference. *Bioinformatics* **35**, 149–151 (2019).
36. Forslund, K., Henricson, A., Hollich, V. & Sonnhammer, E. L. L. Domain tree-based analysis of protein architecture evolution. *Mol. Biol. Evol.* **25**, 254–264 (2008).
37. Forslund, S. K., Kaduk, M. & Sonnhammer, E. L. L. Evolution of Protein Domain Architectures. *Methods Mol. Biol.* **1910**, 469–504 (2019).
38. Haider, C., Kavic, M. & Sonnhammer, E. L. L. TreeDom: a graphical web tool for analysing domain architecture evolution. *Bioinformatics* **32**, 2384–2385 (2016).
39. Adebali, O., Ortega, D. R. & Zhulin, I. B. CDvist: a webserver for identification and visualization of conserved domains in protein sequences. *Bioinformatics* **31**, 1475–1477 (2015).
40. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research* **51**, D445–D451 (2023).
41. Schläpfer, P., Mehta, D., Ridderikhoff, C. & Uhrig, R. G. DomainViz: intuitive visualization of consensus domain distributions across groups of proteins. *Nucleic Acids Research* **49**, W169–W173 (2021).
42. Wang, H. *et al.* VisProDom: an interactive Shiny/R application for displaying protein domains with transcriptional features. *BMC Genomics* **23**, 472 (2022).
43. Del Conte, A. *et al.* CAID prediction portal: a comprehensive service for predicting intrinsic disorder and binding regions in proteins. *Nucleic Acids Research* **51**, W62–W69 (2023).
44. Jiang, A. K., Zhao, J. & Jiang, X. EzSEA: an interactive web interface for enzyme sequence evolution analysis. *Bioinformatics Advances* **5**, vbaf118 (2024).
45. Paysan-Lafosse, T. *et al.* The Pfam protein families database: embracing AI/ML. *Nucleic Acids*

Research **53**, D523–D534 (2025).
46. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2020).
47. RStudio Team. *RStudio: Integrated Development Environment for R*. (RStudio, PBC., Boston, MA, 2020).
48. Allaire, J. J. *et al. Rmarkdown: Dynamic Documents for R*. (2020).
49. JJ Allaire, T., Yihui Xie, R. Foundation, Hadley Wickham, Journal of Statistical Software, RStudio, Ramnath Vaidyanathan, Association for Computing Machinery, Carl Boettiger, Elsevier, Karl Broman, Kirill Mueller, Bastiaan Quast, Randall Pruim, Ben Marwick, Charlotte Wickham, Oliver Keyes, Miao Yu, Daniel Emaasit, Thierry Onkelinx, Alessandro Gasparini, Marc-Andre Desautels, Dominik Leutnant, MDPI & Francis, C. D., Oguzhan Ögreden, Dalton Hance, Daniel Nüst, Petter Uvesten, Elio Campitelli, John Muschelli, Alex Hayes, Zhian N. Kamvar, Noam Ross, Robrecht Cannoodt, Duncan Luguern, David M. Kaplan, Sebastian Kreutzer, Shixiang Wang, Jay Hesselberth, Alfredo Hernández. *Article Formats for R Markdown*. (2020).
50. Xie, Y., Allaire, J. J. & Grolemund, G. *R Markdown: The Definitive Guide*. (Chapman and Hall/CRC, Boca Raton, Florida, 2018).
51. Müller, K. *Here: A Simpler Way to Find Your Files*. (2017).
52. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).
53. Jake Conway, N. G. *A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. (2019).
54. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* **20**, 1983–1992 (2014).
55. Perry, M. *Flexible Heatmaps for Functional Genomics and Sequence Features*. (2016).
56. Fellows, I. *Word Clouds*. (2018).
57. Dawei Lang, G. C. *Create Word Cloud by 'htmlwidget*. (2018).
58. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
59. Enrico Bonatesta, U. B., Christoph Horejs-Kainrath. *Multiple Sequence Alignment*. (2020).
60. Hao ZhuThomas Travison, D. M., Timothy Tsai, Will Beasley, Yihui Xie, GuangChuang Yu, Stéphane Laurent, Rob Shepherd, Yoni Sidi, Brian Salzer, George Gui, Yeliang Fan. *Construct Complex Table with 'kable' and Pipe Syntax*. (2020).
61. Mike Bostock, C. Y., Kerry Rodden, Kevin Warne, Kent Russell, Florian Breitwieser. *Sunburst 'Htmlwidget'*. (2020).
62. Chang, W., Cheng, J., Allaire, J. J., Xie, Y. & McPherson, J. *Shiny: Web Application Framework for R*. (2019).
63. Winston Chang, R. C. T., Joe Cheng, JJ Allaire, Yihui Xie, Jonathan McPherson, RStudio, jQuery Foundation, jQuery contributors, jQuery UI contributors, Mark Otto, Jacob Thornton, Bootstrap contributors, Twitter, Inc, Alexander Farkas, Scott Jehl, Stefan Petre, Andrew Rowls, Dave Gandy, Brian Reavis, Kristopher Michael Kowal, es5-shim contributors, Denis Ineshin, Sami Samhuri, SpryMedia Limited, John Fraser, John Gruber, Ivan Sagalaev. *Web Application Framework for R*. (2020).
64. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36-42 (2013).
65. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).
66. Tatusova, T. *et al.* Update on RefSeq microbial genomes resources. *Nucleic Acids Res.* **43**, D599-605 (2015).
67. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136-143 (2012).
68. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein

sequences. *Protein Sci.* **27**, 135–145 (2018).

69. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419 (2021).
70. Wang, J. *et al.* The conserved domain database in 2023. *Nucleic Acids Research* **51**, D384–D388 (2023).
71. Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Microbial genome analysis: the COG approach. *Brief. Bioinformatics* **20**, 1063–1070 (2019).
72. Geer, L. Y., Domrachev, M., Lipman, D. J. & Bryant, S. H. CDART: protein homology by domain architecture. *Genome Res.* **12**, 1619–1623 (2002).
73. Nielsen, H. Predicting Secretory Proteins with SignalP. *Methods Mol. Biol.* **1611**, 59–73 (2017).
74. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).
75. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.* **35**, W429-432 (2007).
76. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
77. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244-248 (2005).
78. Cole, C., Barber, J. D. & Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **36**, W197-201 (2008).
79. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
80. Yeats, C., Lees, J., Carter, P., Sillitoe, I. & Orengo, C. The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. *Nucleic Acids Res* **39**, W546-550 (2011).
81. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research* **51**, D690–D699 (2023).
82. Ghosh, A. *et al.* From sequence to signature: Machine learning uncovers multiscale feature landscapes that predict AMR across ESKAPE pathogens. Preprint at https://doi.org/10.1101/2025.07.03.663053 (2025).