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

Jacob D Krol[1,#], Joseph T Burke[2,3,#], Samuel Z Chen[3,4,#], Lo Sosinski[3,5,#], Faisal S Alquaddoomi[1], Evan P Brenner[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 Medical Campus, 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 Medical Campus, 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 conserved features, lineage-specific variants, and their potential functions. *MolEvolvR* (https://jravilab.org/molevolvr) is a novel web-app enabling researchers to visualize the molecular evolution of their proteins of interest in a phylogenetic context across the tree of life, spanning all superkingdoms. The web-app 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). Thus, *MolEvolvR* is a powerful, easy-to-use web interface for computational protein characterization.

Keywords

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

The rate of protein family discovery 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, antibiotic resistance, or stress response. Identifying and characterizing the complete molecular systems involved in such processes requires functional knowledge of the proteins involved. Many studies[2–16] have demonstrated the power of phylogenetic and molecular evolutionary analysis in determining the molecular functions of such proteins. A variety of individual tools[17–41] exist for protein sequence similarity searches or ortholog detection, delineating co-occurring domains (domain architectures), and building multiple sequence alignments and phylogenetic trees. However, despite providing complementary insight, these tools are disjointed and often require both technical skill and background knowledge for meaningful use. A unified, user-friendly web-based framework is needed for biologists to functionally characterize their unknown/understudied protein(s) of interest by exhaustively identifying relevant homologs with shared motifs, domains, and domain architectures. Further, this data spans multiple super-kingdoms and phyletic scales whose interpretable integration requires effective summarization and visualization to provide functional insights.

Here, we present *MolEvolvR*, a web application that comprehensively characterizes proteins in a streamlined, easy-to-use platform (**Fig. 1**; accessible at jravilab.org/molevolvr). *MolEvolvR* performs homology searches across the tree of life and reconstructs domain architecture by functionally characterizing the input proteins and each of their homologs, presenting these results in the context of evolution. The computational evolutionary framework underlying *MolEvolvR* is written using custom R[42–57] and shell scripts. The web application is built with an R/Shiny[44,58,59] framework with additional user interface customizations in HTML, Javascript, and CSS, and has been tested on Chrome, Brave, Firefox, and Safari browsers on Mac, Windows, and Linux operating systems.

Consider a researcher starting with a protein of unknown function, either falling out of a genetic screen for a phenotype or an uncharacterized protein in a genomic context of interest. In Step 1, *MolEvolvR* resolves this query protein into its constituent domains and uses each domain for iterative homology searches[18,19] across the tree of life[60–63]. This divide-and-conquer strategy is domain-sensitive by design and captures remote homologs missed by homology searches using the full-length protein sequences alone. In Step 2, to delineate molecular function, *MolEvolvR* reconstructs the domain architecture by characterizing the protein and each of its homologs, combining: 1) sequence alignment and clustering algorithms for domain detection[17,22,23,64]; 2) profile matching against protein domain and orthology databases[30,31,65–68]; and 3) prediction algorithms for signal peptides[31,69], transmembrane regions[31,70–72], cellular localization[31,71], and secondary/tertiary structures[31,73–76]. This analysis illustrates the breadth of MolEvolvR's analytic capability: results entail a detailed molecular characterization of the initial query protein(s) and of their homologs integrated across the superkingdoms of life, along with any lineage-specific functional variants.

MolEvolvR is versatile, accommodating a variety of inputs, investigating wide-ranging questions, and producing diverse potential outputs. 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 features restricted to certain pathogenic groups to discover virulence factors/diagnostic targets [**Fig. 1C**]. 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 for each result, *MolEvolvR* also provides graphical summaries that combine these results in the context of evolution: i) structure-based multiple sequence alignments and phylogenetic trees; ii) domain proximity networks to consolidate results from all co-occurring domains (across homolog domain architectures); 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 web-app contains detailed documentation about all these options, including case studies (see *Supplementary Material*) and frequently asked questions (FAQs).

A specific instance of the web-app applied to study several PSP stress response proteins (present across the tree of life) can be found here: https://jravilab.shinyapps.io/psp_evolution[2]. *MolEvolvR* is a generalized web-server of this web-app. To demonstrate its broad applicability, we have applied the approach underlying *MolEvolvR* to study several systems, including proteins/operons in zoonotic pathogens, *e.g.*, nutrient acquisition systems in *Staphylococcus aureus*[4,5], novel phage defense system in *Vibrio cholerae*[6], surface layer proteins in *Bacillus anthracis*[7], helicase operators in bacteria[8], 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.

Thus, *MolEvolvR* (jravilab.org/molevolvr) is a flexible, user-friendly, and powerful interactive web tool that allows researchers of any level of bioinformatics experience to bring molecular evolution and phylogeny to bear on their proteins of interest.

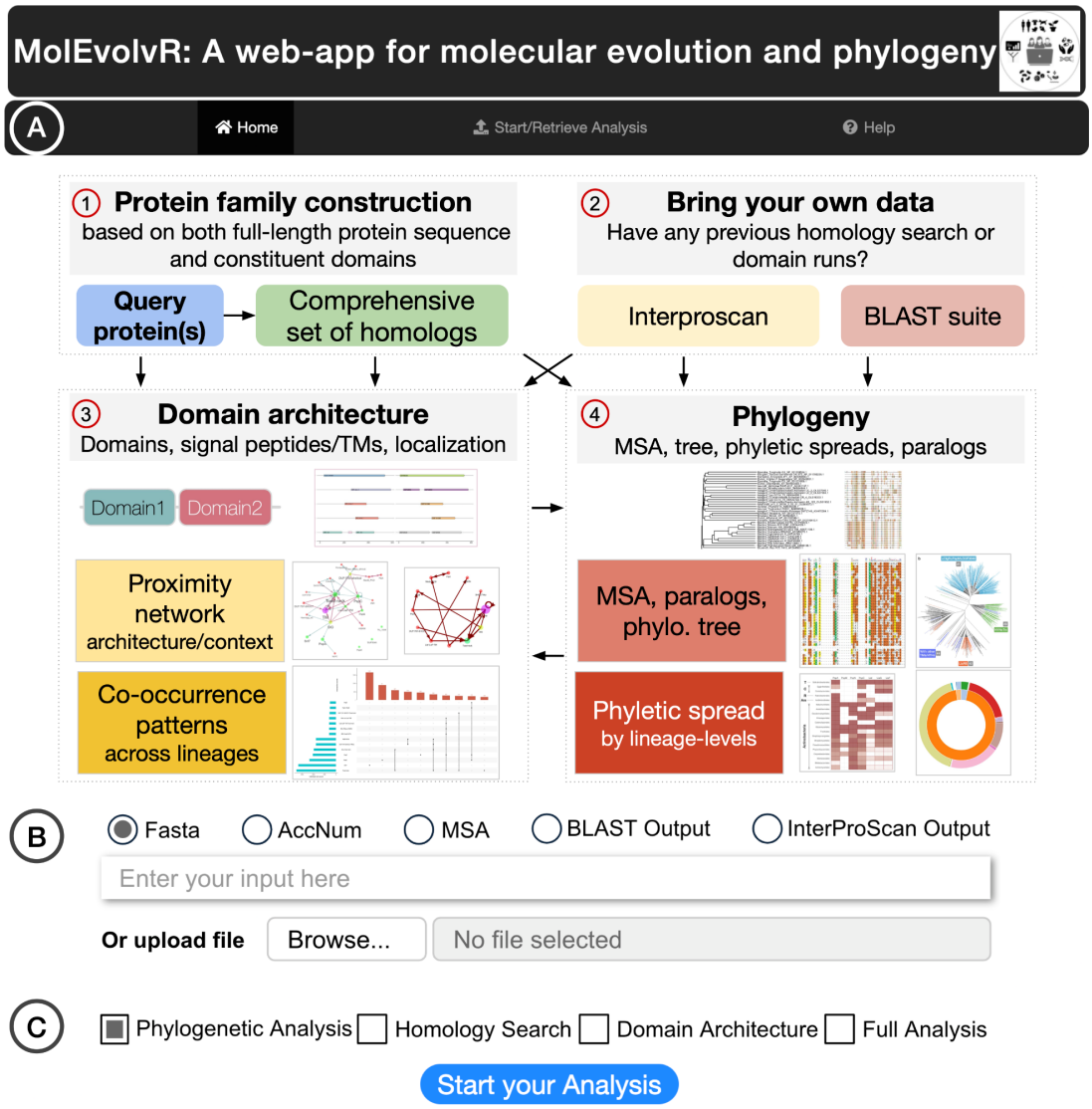

Figure 1. Overview of *MolEvolvR*. A. *MolEvolvR* allows users to start with protein(s) of interest and perform the full analysis (1+3+4), only protein characterization (1+3), or only homology searches (1+4); or start with external outputs from BLAST or Interproscan for further analysis, summarization, and visualization (2+3+4). *MolEvolvR* is interactive, queryable, and customizable. **B.** Multiple input options in *MolEvolvR*: FASTA, NCBI accession numbers, pre-computed 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**.

Acknowledgments

We would like to thank members of the JRaviLab (Elliot Majlessi, Karn Jongnarangsin, Kewalin Samart, Dave Bunten) for testing the web-app at various stages and providing the authors with several iterations of constructive feedback. We are also extremely grateful to Krishnan Raghunathan, Premal Shah, Arjun Krishnan, Emily Meyer, Jerome McKay, Dave Bunten, and members of the Krishnan lab (Kayla Johnson, Nathaniel Hawkins) for 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, thanks to collaborations with L Aravind, Neal Hammer, Christopher Waters, Antonella Fioravanti, Jonathan Hardy, Kayla Conner, Christina Stallings, Helen Blaine, Nina Wale, Victor DiRita, and Stephanie Shames.

Funding

We would like to thank our funding sources: Endowed Research Funds from the College of Veterinary Medicine, Michigan State University, NSF-funded BEACON funding support, and the University of Colorado Anschutz start-up funds 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, JDK, SZC, LS, and JR acquired the data, performed all the analyses, and made the figures and tables, specifically JTB, JDK, SZC, LS, JR wrote the backend code, JTB, SZC, JDK, FSA, VPR, JR built the R/Shiny web-app, JBJ, KMR, SA, and FSA set up the server that hosts the web-app; JTB, JDK, FSA, EPB, VPR, EPW, JR tested the web-app rigorously; JTB, JDK, LS, EPB, EPW, JR worked on the help documentation; LS and JR wrote the first draft of the manuscript; JTB, JDK, LS, EPB, 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. *bioRxiv [Preprint]* 2020.09.24.301986 (2020) doi:10.1101/2020.09.24.301986.
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* (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. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

18. Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7**, 12 (2012).
19. 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).
20. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
21. 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).
22. Magis, C. *et al.* T-Coffee: Tree-based consistency objective function for alignment evaluation. *Methods Mol. Biol.* **1079**, 117–129 (2014).
23. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
24. Lassmann, T. Kalign 3: multiple sequence alignment of large data sets. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz795.
25. 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.* (2020) doi:10.1093/nar/gkaa243.
26. Klimchuk, O. I. *et al.* COGNAT: a web server for comparative analysis of genomic neighborhoods. *Biol. Direct* **12**, 26 (2017).
27. Shmakov, S. A. *et al.* Systematic prediction of functionally linked genes in bacterial and archaeal genomes. *Nat Protoc* **14**, 3013–3031 (2019).
28. 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).
29. Persson, E., Kaduk, M., Forslund, S. K. & Sonnhammer, E. L. L. Domainoid: domain-oriented orthology inference. *BMC Bioinformatics* **20**, 523 (2019).
30. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344–D354 (2021).
31. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
32. 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).
33. Cosentino, S. & Iwasaki, W. SonicParanoid: fast, accurate and easy orthology inference. *Bioinformatics* **35**, 149–151 (2019).
34. 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).
35. Forslund, S. K., Kaduk, M. & Sonnhammer, E. L. L. Evolution of Protein Domain Architectures. *Methods Mol. Biol.* **1910**, 469–504 (2019).
36. Haider, C., Kavic, M. & Sonnhammer, E. L. L. TreeDom: a graphical web tool for analysing domain architecture evolution. *Bioinformatics* **32**, 2384–2385 (2016).
37. 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).

38. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research* **51**, D445–D451 (2023).
39. 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).
40. Wang, H. *et al.* VisProDom: an interactive Shiny/R application for displaying protein domains with transcriptional features. *BMC Genomics* **23**, 472 (2022).
41. 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).
42. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2020).
43. RStudio Team. *RStudio: Integrated Development Environment for R*. (RStudio, PBC., 2020).
44. Allaire, J. J. *et al. rmarkdown: Dynamic Documents for R*. (2020).
45. 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).
46. Xie, Y., Allaire, J. J. & Grolemund, G. *R Markdown: The Definitive Guide*. (Chapman and Hall/CRC, 2018).
47. Müller, K. *here: A Simpler Way to Find Your Files*. (2017).
48. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).
49. Jake Conway, N. G. *A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. (2019).
50. 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).
51. Perry, M. *Flexible Heatmaps for Functional Genomics and Sequence Features*. (2016).
52. Fellows, I. *Word Clouds*. (2018).
53. Dawei Lang, G. C. *Create Word Cloud by 'htmlwidget*. (2018).
54. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
55. Enrico Bonatesta, U. B., Christoph Horejs-Kainrath. *Multiple Sequence Alignment*. (2020).
56. 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).
57. Mike Bostock, C. Y., Kerry Rodden, Kevin Warne, Kent Russell, Florian Breitwieser. *Sunburst 'Htmlwidget'*. (2020).

58. Chang, W., Cheng, J., Allaire, J. J., Xie, Y. & McPherson, J. *shiny: Web Application Framework for R*. (2019).
59. 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).
60. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36-42 (2013).
61. 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).
62. Tatusova, T. *et al.* Update on RefSeq microbial genomes resources. *Nucleic Acids Res.* **43**, D599-605 (2015).
63. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136-143 (2012).
64. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
65. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419 (2021).
66. Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).
67. 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).
68. Geer, L. Y., Domrachev, M., Lipman, D. J. & Bryant, S. H. CDART: protein homology by domain architecture. *Genome Res.* **12**, 1619–1623 (2002).
69. Nielsen, H. Predicting Secretory Proteins with SignalP. *Methods Mol. Biol.* **1611**, 59–73 (2017).
70. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).
71. 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).
72. 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).
73. 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).
74. Cole, C., Barber, J. D. & Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **36**, W197-201 (2008).
75. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
76. 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).