ASM | Microbiology Resource Announcements

Check for updates

| Bacteriology | Announcement

DiMER: a modular framework for replacing uninformative annotations in protein datasets

Ellis L. Torrance,[1] Kelly P. Williams,[1] Joseph S. Schoeniger,[2] Catherine M. Mageeney[3]

AUTHOR AFFILIATIONS See affiliation list on p. 3.

ABSTRACT DiMER (https://github.com/sandialabs/DiMER) is a modular Python framework using user-provided BLAST and HMM databases to assign annotations while avoiding placeholders (e.g., hypothetical protein). DiMER was applied to a curated list of gold standard mobile genetic elements and increased functional annotations by 24% (49 to 73%).

KEYWORDS annotation, microbes, phage, mobile elements

Accurate and informative functional annotation remains a central challenge in microbial genomics, paramount to understanding a microbe's biological role in a community, identification of novel synthetic biology tools, and enhanced metabolomic and interaction modeling (1). This challenge is particularly difficult for mobile genetic elements (MGEs), such as prophages and integrative conjugative elements, since they contain large amounts of "dark matter" (1). Homology-based annotation is standard practice, but many microbial sequences still retain large fractions of coding sequences labeled with uninformative annotations (termed placeholders; e.g., "hypothetical protein" or "uncharacterized protein") (1). These labels persist, even when more informative homolog-derived descriptions exist, because many annotation workflows select only the single best-scoring hit per query and do not explicitly discriminate against placeholder terms. Consequently, downstream comparative analyses and functional interpretation of MGE gene inventories are hindered, and placeholder terms are propagated in derived databases and training sets.

We developed DiMER, a modular Python framework intended to quickly reduce placeholder annotation retention by explicitly incorporating user knowledge of their genomic system to identify uninformative annotations. DiMER takes a protein FASTA file (with or without existing annotations) and queries user-specified databases in HMMER or DIAMOND format (2, 3). For each query protein, DiMER records all best matches to the queried databases in a single table and selects a single final annotation using e-value ranking under a user-defined threshold (default: 1e−10) (Fig. 1A). During this selection, DiMER preferentially avoids any user-defined placeholder annotations (listed in placeholders.txt) and retains the best available hit.

DiMER can run in "gentle" mode where it will replace only the initial gene annotations that were identified as placeholders (via placeholders.txt) or in "force" mode, which replaces all annotations based on e-value across the queried databases while still preferentially avoiding placeholders (via placeholders.txt) (Fig. 1A). DiMER supports iterative refinement where users can update the uninformative annotation list based on annotations encountered in the DiMER output (e.g., "DUF#" or "ORF#"). DiMER can be rerun to improve annotation without repeating expensive database searches when prior results are present. DiMER was built using Python 3 and requires HMMER v3.1b2 (2), Diamond v2.1.10 (3), and the Python library Pandas to run. Installation and running are straightforward, and instructions are available at https://github.com/sandialabs/DiMER.

Editor Patrick D. Schloss, University of Michigan, Ann Arbor, Michigan, USA

Address correspondence to Catherine M. Mageeney, cmmagee@sandia.gov.

The authors declare no conflict of interest.

See the funding table on p. 3.

Ed Note: This article was handled by Pat Schloss, who acted as a guest editor in consultation with the editor in chief, Irene Newton. Dr. Schloss's role as the guest editor of this article is outside of his duties as the chair of the Journals Committee.

Received 29 June 2026
Accepted 20 August 2026
Published 31 August 2026

Copyright © 2026 Torrance et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

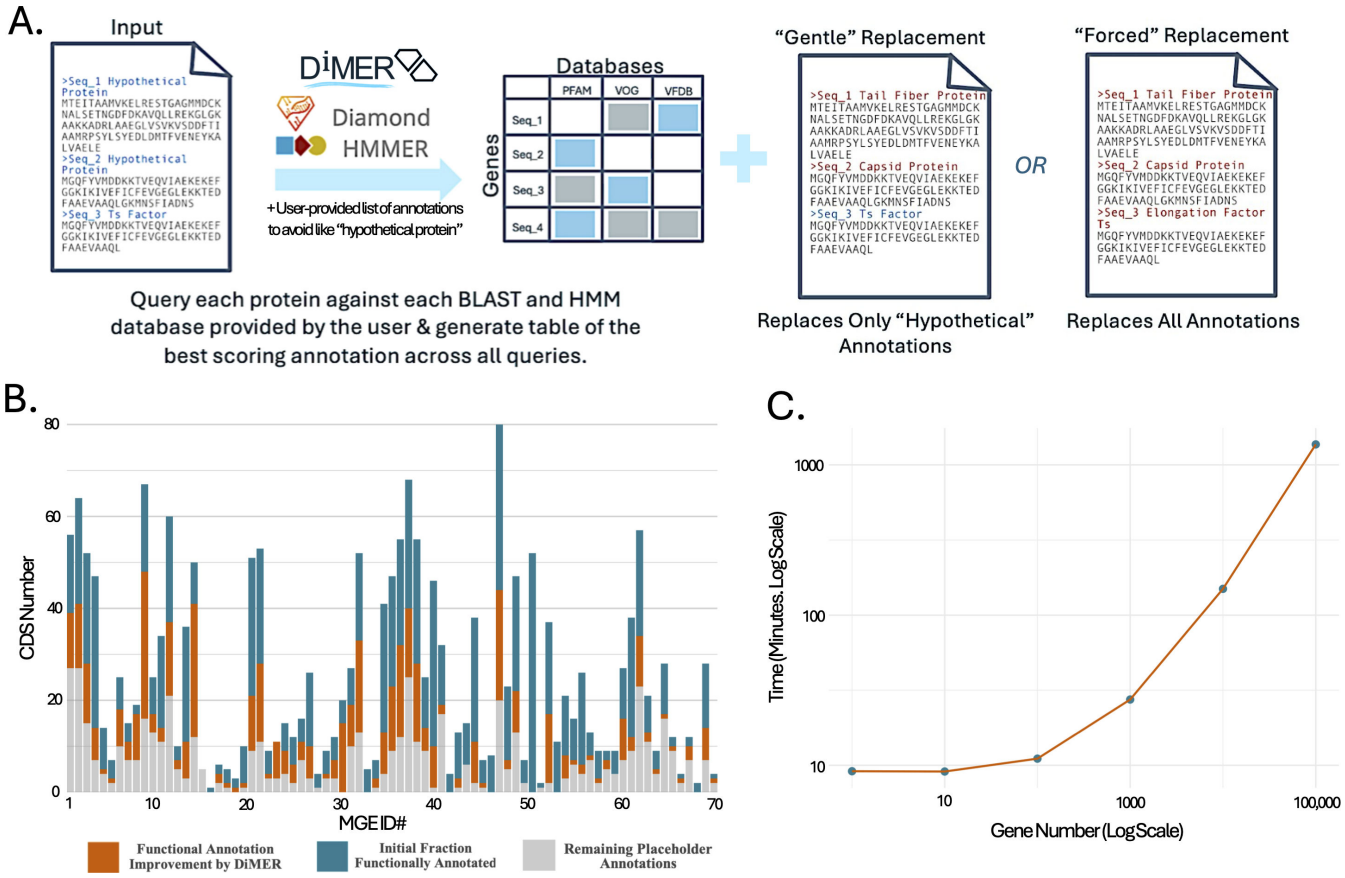


FIG 1 Overview and performance metrics of DiMER. (A) DiMER queries user-provided databases to annotate proteins while selectively avoiding annotations deemed as placeholders whenever possible. (B) Stacked bar plot demonstrating replacement of placeholder annotations in bacterial MGEs (n = 79) using DiMER. (C) Plot demonstrating DiMER runtime in minutes across a randomly generated data set of differing numbers of genes in log scale.

As a demonstration, DiMER was evaluated on experimentally validated MGEs (n = 79) (4) downloaded from the National Center for Biotechnology Information (NCBI) that contained 1,998 predicted ORFs. Of these, 1,021 (51%) were labeled with placeholder terms by NCBI/PGAP at download. We compiled 15 MGE-relevant databases (three DIAMOND-formatted protein databases and 12 HMM profile databases [5–20]) and ran DiMER in gentle mode, replacing 483 placeholder annotations (47% of the placeholder fraction) and increasing the functionally annotated proportion from 49 to 73% (Fig. 1B). Furthermore, with these databases, DiMER required only ~10 min to run on genetic data sets with <1,000 genes (Fig. 1C). These results demonstrate the utility of DiMER to improve MGE annotation. Furthermore, this software can be easily implemented into existing annotation workflows for any organism of interest and reduces time searching across multiple databases for the most informative and functionally accurate annotations.

ACKNOWLEDGMENTS

We thank Raga Krishnakumar for thoughtful feedback on the manuscript. We additionally thank Elriana Styles, Joanna Steczynska, and Jessica Trinh for initial software testing and user feedback.

Research funding was provided by the U.S. Department of Energy, Office of Science through the Genomic Science Program, Office of Biological and Environmental Research under the Secure Biosystems Design Science Focus Area, Intrinsic Control for Genome and Transcriptome Editing in Communities (InCoGenTEC).

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all rights, titles, and interests in and to the article and is solely responsible for its contents. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this article or allow others to do so for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan https://www.energy.gov/downloads/doe-public-access-plan.

AUTHOR AFFILIATIONS

[1]Systems Biology Department, Sandia National Laboratories, Livermore, California, USA
[2]Biological, Radiation, and Signature Science, Technology, & Engineering Center, Sandia National Laboratories, Livermore, California, USA
[3]Biotechnology & Bioengineering Department, Sandia National Laboratories, Livermore, California, USA

AUTHOR ORCIDs

Kelly P. Williams http://orcid.org/0000-0002-2606-9562
Joseph S. Schoeniger http://orcid.org/0000-0002-2719-7435
Catherine M. Mageeney http://orcid.org/0000-0001-7969-1622

FUNDING

Funder	Grant(s)	Author(s)
Biological and Environmental Research	24-026159	Joseph S. Schoeniger

AUTHOR CONTRIBUTIONS

Ellis L. Torrance, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review and editing | Kelly P. Williams, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Writing – original draft, Writing – review and editing | Joseph S. Schoeniger, Funding acquisition, Supervision, Validation, Writing – review and editing | Catherine M. Mageeney, Conceptualization, Supervision, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

The dataset of MGE analyzed in this publication was taken from Langille et al. (2008) (4) and downloaded from NCBI by accession number (listed at: https://doi.org/10.5281/zenodo.15724972). The software DiMER and the associated databases used in this analysis are available at github.com/sandialabs/DiMER under the CC0-1.0 license. A copy with an associate DOI is available at https://doi.org/10.5281/zenodo.21893019.

REFERENCES

1. Escudeiro P, Henry CS, Dias RPM. 2022. Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. Curr Res Microb Sci 3:100159. https://doi.org/10.1016/j.crmicr.2022.100159
2. Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960. https://doi.org/10.1093/bioinformatics/bti125
3. Buchfink B, Reuter K, Drost HG. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. https://doi.org/10.1038/s41592-021-01101-x
4. Langille MGI, Hsiao WWL, Brinkman FSL. 2008. Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics 9:329. https://doi.org/10.1186/1471-2105-9-329
5. Liu B, Zheng D, Zhou S, Chen L, Yang J. 2022. VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res 50:D912–D917. https://doi.org/10.1093/nar/gkab1107
6. Grazziotin AL, Koonin EV, Kristensen DM. 2017. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res 45:D491–D498. https://doi.org/10.1093/nar/gkw975
7. Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res 31:371–373. https://doi.org/10.1093/nar/gkg128
8. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419. https://doi.org/10.1093/nar/gkaa913
9. Payne LJ, Todeschini TC, Wu Y, Perry BJ, Ronson CW, Fineran PC, Nobrega FL, Jackson SA. 2021. Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. Nucleic Acids Res 49:10868–10878. https://doi.org/10.1093/nar/gkab883
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
11. Bertelli C, Brinkman FSL. 2018. Improved genomic island predictions with IslandPath-DIMOB. Bioinformatics 34:2161–2167. https://doi.org/10.1093/bioinformatics/bty095
12. Brown CL, Mullet J, Hindi F, Stoll JE, Gupta S, Choi M, Keenum I, Vikesland P, Pruden A, Zhang L. 2022. mobileOG-db: a manually curated database of protein families mediating the life cycle of bacterial mobile genetic elements. Appl Environ Microbiol 88:e0099122. https://doi.org/10.1128/aem.00991-22
13. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 37:D233–D238. https://doi.org/10.1093/nar/gkn663
14. Cury J, Abby SS, Doppelt-Azeroual O, Néron B, Rocha EPC. 2020. Identifying conjugative plasmids and integrative conjugative elements with CONJscan. Methods Mol Biol 2075:265–283. https://doi.org/10.1007/978-1-4939-9877-7_19
15. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, Tyson GH, Klimke W. 2021. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 11:12728. https://doi.org/10.1038/s41598-021-91456-0
16. Galperin MY, Vera Alvarez R, Karamycheva S, Makarova KS, Wolf YI, Landsman D, Koonin EV. 2025. COG database update 2024. Nucleic Acids Res 53:D356–D363. https://doi.org/10.1093/nar/gkae983
17. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, Toussaint A, Petit M-A, Enault F. 2021. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform 3:lqab067. https://doi.org/10.1093/nargab/lqab067
18. Trgovec-Greif L, Hellinger H-J, Mainguy J, Pfundner A, Frishman D, Kiening M, Webster NS, Laffy PW, Feichtinger M, Rattei T. 2024. VOGDB-database of virus orthologous groups. Viruses 16:1191. https://doi.org/10.3390/v16081191
19. Wang M, Liu G, Liu M, Tai C, Deng Z, Song J, Ou HY. 2024. ICEberg 3.0: functional categorization and analysis of the integrative and conjugative elements in bacteria. Nucleic Acids Res 52:D732–D737. https://doi.org/10.1093/nar/gkad935
20. Zhang KY, Gao YZ, Du MZ, Liu S, Dong C, Guo FB. 2019. Vgas: a viral genome annotation system. Front Microbiol 10:184. https://doi.org/10.3389/fmicb.2019.00184