Received: 30 December 2024 | Revised: 7 May 2025 | Accepted: 9 May 2025

DOI: 10.1002/pro.70177

TOOLS FOR PROTEIN SCIENCE

AcrDB update: Predicted 3D structures of anti-CRISPRs in human gut viromes

Minal Khatri | N. R. Siva Shanmugam | Xinpeng Zhang | Revanth Sai Kumar Reddy Patel | Yanbin Yin

Nebraska Food for Health Center, Department of Food Science and Technology, University of Nebraska—Lincoln, Lincoln, Nebraska, USA

Correspondence
Yanbin Yin, Nebraska Food for Health Center, Department of Food Science and Technology, University of Nebraska—Lincoln, Lincoln, NE 68588, USA.
Email: yyin@unl.edu

Funding information
National Institutes of Health, Grant/Award Numbers: R01GM140370, R21AI171952; U. S. Department of Agriculture (USDA), Grant/Award Number: 58-8042-9-089

Review Editor: Nir Ben-Tal

Abstract

Anti-CRISPR (Acr) proteins play a key role in phage-host interactions and hold great promise for advancing genome-editing technologies. However, finding new Acrs has been challenging due to their low sequence similarity. Recent advances in protein structure prediction have opened new pathways for Acr discovery by using 3D structure similarity. This study presents an updated AcrDB, with the following new features not available in other databases: (1) predicted Acrs from human gut virome databases, (2) Acr structures predicted by AlphaFold2, (3) a structural similarity search function to allow users to submit new sequences and structures to search against 3D structures of experimentally known Acrs. The updated AcrDB contains predicted 3D structures of 795 candidate Acrs with structural similarity (TM-score ≥0.7) to known Acrs supported by at least two of the three non-sequence similarity-based tools (TM-Vec, Foldseek, AcrPred). Among these candidate Acrs, 121 are supported by all three tools. AcrDB also includes 3D structures of 122 experimentally characterized Acr proteins. The 121 most confident candidate Acrs were combined with the 122 known Acrs and clustered into 163 sequence similarity-based Acr families. The 163 families were further subject to a structure similarity-based hierarchical clustering, revealing structural similarity between 44 candidate Acr (cAcr) families and 119 known Acr families. The bacterial hosts of these 163 Acr families are mainly from *Bacillota*, *Pseudomonadota*, and *Bacteroidota*, which are all dominant gut bacterial phyla. Many of these 163 Acr families are also co-localized in Acr operons. All the data and visualization are provided on our website: https://pro.unl.edu/AcrDB.

KEYWORDS
anti-CRISPR, anti-defense genes, bacterial immunity, CRISPR-Cas, structure similarity, virome

1 | INTRODUCTION

Anti-CRISPR proteins (Acrs) were first discovered in 2013 in *Pseudomonas* phages and prophages (Bondy-Denomy et al., 2013). Most Acr proteins are short (<200 amino acids) and produced by phages and other mobile genetic elements to inhibit the CRISPR-Cas immune systems of their prokaryotic hosts. Therefore, Acrs can turn off host CRISPR-Cas system acting as a "naturally occurring off-switch" for these systems. Consequently, Acrs have great potential to serve as regulators or modulators of CRISPR-Cas genome editing tools for safer and more controllable genome engineering. Together with many other recently

Minal Khatri and N. R. Siva Shanmugam are co-first authors.

discovered anti-defense systems (Tesson et al., 2025; Yan et al., 2023), Acrs are key players of the endless arms race between mobile genetic elements like phages and their prokaryote hosts (Mayo-Muñoz et al., 2024), ensuring their co-existence and co-evolution in a well-balanced microbial ecosystem.

Since 2013, 122 experimentally characterized Acr proteins that inhibit 13 CRISPR-Cas systems have been published. Many of these Acrs were identified with the help of bioinformatics (Bondy-Denomy et al., 2018; Pawluk et al., 2018; Stanley & Maxwell, 2018). Notably, very few sequence similarities are found among these Acrs, and most of them do not have any conserved Pfam domains (Yang et al., 2022). Therefore, sequence similarity search is of limited use in the discovery of new Acrs (Makarova et al., 2023). To address this limitation, several computational tools leveraging machine learning and deep learning have been developed to predict potential Acrs on the genome scale, which include AcRanker (Eitzinger et al., 2020), PaCRISPR (Wang et al., 2020), AcrPred (Dao et al., 2023), AcrNET (Li et al., 2023), and DeepAcr (Wandera et al., 2022) as summarized in Table 1. In addition, databases for known and predicted Acrs are also available, which include anti-CRISPRDB (Dong et al., 2022), CRISPRminer (Zhang et al., 2018), AcrCatalog (Gussow et al., 2020), AcrHub (Wang et al., 2021), and AcrDB (Huang et al., 2021) (Table 1), providing valuable resources for ongoing research and discovery.

Acr encoding genes often form operons with putative transcription regulator genes that encode Acr-associated (Aca) proteins (Bondy-Denomy, 2018; Borges et al., 2017), which negatively regulate Acr expression (Birkholz et al., 2024). We have focused on developing new bioinformatics tools for identifying *Acr-Aca* operons (Yang et al., 2022; Yin et al., 2019). We published AcrFinder (Yi et al., 2020) as a web server in 2020 to scan genomes for *Acr-Aca* operons. In 2021, we further developed AcrDB (Huang et al., 2021) (https://bcb.unl.edu/AcrDB/) to include *Acr-Aca* operons predicted from over 19,000 RefSeq prokaryotic and viral genomes predicted by AcrFinder, AcRanker, and PaCRISPR. However, not all Acrs are associated

TABLE 1 Online bioinformatics tools for Acr research.

Name (reference)	Year	Resource provided	Features
CRISPRminer (Zhang et al., 2018)	2018	Database	Experimentally characterized Acrs and their homologs and genomic context
Acr nomenclature (Bondy-Denomy et al., 2018)	2018	Google spreadsheets	Experimentally characterized Acrs and Aca nomenclature
AcrCatalog (Gussow et al., 2020)	2020	Database and model code	Predicted Acrs from decision tree ML classifier + heuristic filtering
AcRanker (Eitzinger et al., 2020)	2020	Web server and standalone	XGBoost classifier using AA biases
AcrFinder (Yi et al., 2020)	2020	Web server and standalone	Workflow combining Homology + GBA + self-targeting and user-friendly website
PaCRISPR (Wang et al., 2020)	2020	Web server	SVM classifier using PSSMs to capture evolutionary features, Decision tree classifier using six sequence features
AcrDB (Huang et al., 2021)	2021	Database and Web Server	Predicted Acr-Aca operons from 19,000 RefSeq genomes identified with AcrFinder and refined by AcRanker and PaCRISPR along with their homologs and analysis of their gene neighborhood.
AcrHub (Wang et al., 2021)	2021	Database and Web Server	Experimentally validated and predicted Acrs identified using three tools, PaCRISPR, AcRanker and HMM predictor and their gene neighborhood
Anti-CRISPRdb v2.2 (Dong et al., 2022)	2022	Database	Experimentally characterized Acrs and their homologs from PSI-BLAST search. Inhibitory mechanisms and neighbors of curated anti-CRISPR proteins
AcaFinder (Yang et al., 2022)	2022	Web server and standalone	Identify Aca genes from prokaryotic and phage genomes, using Homology + GBA + Self-targeting
DeepAcr (Wandera et al., 2022)	2022	Standalone	CNN and DNN-based model that integrates evolutionary, structural and sequence features along with transformer-derived (ESM-1b) protein sequence features.
AcrPred (Dao et al., 2023)	2023	Web server and model code	Ensemble model with SVM classifiers using six distinct sequence and evolutionary-based features
AOminer (Yang et al., 2023)	2023	Web server and standalone	Two-state HMM to learn the conserved genomic context of operons that contain known Acr genes or their homologs and distinguish Acr operons and non-Acr operons.
AcrNET (Li et al., 2023)	2023	Webserver	CNNs and fully connected networks (FCNs) using sequence, evolutionary, structural and transformer features

with an Aca gene (Yin et al., 2019) and Acr genes can also co-localize with conserved phage genes such as those encoding capsid, terminase, lysozyme, tail, helicase proteins, and with functionally unknown genes in the gene neighborhood (Pawluk et al., 2018). To fully exploit the genomic context of Acr genes, we developed AOminer (Yang et al., 2023), a machine learning-based tool that uses a two-state Hidden Markov Model (HMM) to identify Acr operons (AOs) by analyzing the conserved genomic context of known Acr genes.

Over the years, at least 40 out of the 122 known Acr proteins have had their PDB structures determined (Sahakyan et al., 2023) often together with their targeted Cas enzymes as a protein complex. These experimentally solved structures (e.g., by cryo-EM) adopt various structural topologies such as all α helices, all β sheets, mixed α/β, and separated $\alpha + \beta$ folds (Davidson et al., 2020). Despite the little structural similarity among them, interestingly, some Acrs share structural similarities with the defense systems in their hosts, for example, toxin-antitoxin and DNA damage-inducible SOS response systems (Sahakyan et al., 2023). Meanwhile, protein structure prediction methods have advanced significantly, particularly with the integration of artificial intelligence for sequence-based structure predictions. Recent state-of-the-art models like AlphaFold2 (Jumper et al., 2021), ESMfold (Lin et al., 2023), and RosettaFold (Baek et al., 2021) have made it possible to predict highly reliable structures across a wide range of protein sequences. This progress has opened new directions by exploring the structures of Acr proteins, whose structures have remained largely unknown. As pointed out in a recent review paper (Makarova et al., 2023), protein 3D structures of Acrs are more conserved than sequences, and a new direction for Acr research is to use structural similarity to facilitate new Acr discovery. Notably, two research papers published in 2023 have predicted 3D structures for experimentally characterized Acrs (Park et al., 2022; Sahakyan et al., 2023) using AlphaFold2 (Jumper et al., 2021). A very recent study modeled 3D structures of 285,00 phage proteins from viromes using AlphaFold2 and compared them to experimentally determined Acr structures (Duan et al., 2024) for new Acrs. This study also included a comprehensive structural comparison of over 100 experimentally validated Acrs, revealing significant structural similarity across Acr families that share little sequence similarity.

Recent studies have introduced a promising direction for structure-guided discovery of Acrs. This motivates an update of AcrDB to incorporate predicted 3D structures of Acrs from human viromes. Additionally, since AcrDB's initial release, the number of experimentally characterized Acrs has grown from 65 to 122. In this article, we present the expanded AcrDB by including Acr operons (AOs) in human viromes. More importantly, we focused on predicted 3D structures for Acrs in these AOs. In our approach, we first predict AOs likely to contain Acrs using AOminer. Proteins from these putative AOs were subjected to 3D structure predictions with ESMfold (Lin et al., 2023) and AlphaFold (Jumper et al., 2021). We further filtered them by comparing against the 122 known Acrs using TM-Vec (Hamamsy et al., 2023), Foldseek (Van Kempen et al., 2024) and the latest machine learning-based tool, AcrPred (Dao et al., 2023).

The updated AcrDB has the following new data and feature compared to the last version: (i) Predicted Acrs from three human gut virome databases GPD (Camarillo-Guerrero et al., 2021), GVD (Gregory et al., 2020), MGV (Gregory et al., 2020), and phage isolate genome database INPHARED (Cook et al., 2021); (ii) 3D structures of 122 experimentally characterized Acr proteins reported in recent papers (Park et al., 2022; Sahakyan et al., 2023) and predicted by ourselves; (iii) AlphaFold2 predicted 3D structures of 795 candidate Acrs with structural similarity (template modeling score or TM-score ≥ 0.7) to known Acrs supported by at least two of the three tools (TM-Vec, Foldseek, or AcrPred); (iv) A structural similarity search function to allow users submit new sequences and structures to search against 3D structures of the 122 known Acrs.

2 | MATERIALS AND METHODS

2.1 | Data source

We focused on three human gut virome databases that contain metagenome-assembled genomes (MAGs): Gut Virome Database (GVD) (Gregory et al., 2020), Gut Phage Database (GPD) (Camarillo-Guerrero et al., 2021), and Metagenomic Gut Virus database (MGV) (Nayfach et al., 2021). We also included a phage isolate genome database INPHARED (Cook et al., 2021). GVD contains 33,432 viral MAGs from 2697 human gut metagenomes, GPD comprises 142,809 MAGs from 28,060 human gut metagenomes, MGV contains 189,680 MAGs from 11,810 human stool metagenomes, and INPHARED has 25,596 complete phage genomes from GenBank.

The overall workflow of data collection and analysis is provided in Figure 1 consisting of the following steps.

2.2 | Anti-CRISPR operon prediction

To identify potential Anti-CRISPR operons (AOs) in the four databases, we used AOminer (Yang et al., 2023). For a more accurate phage-specific gene annotation, we replaced AOminer's default gene prediction tool Prodigal (Hyatt et al., 2010) with Pharokka (Bouras et al., 2023), a recent tool specifically developed for

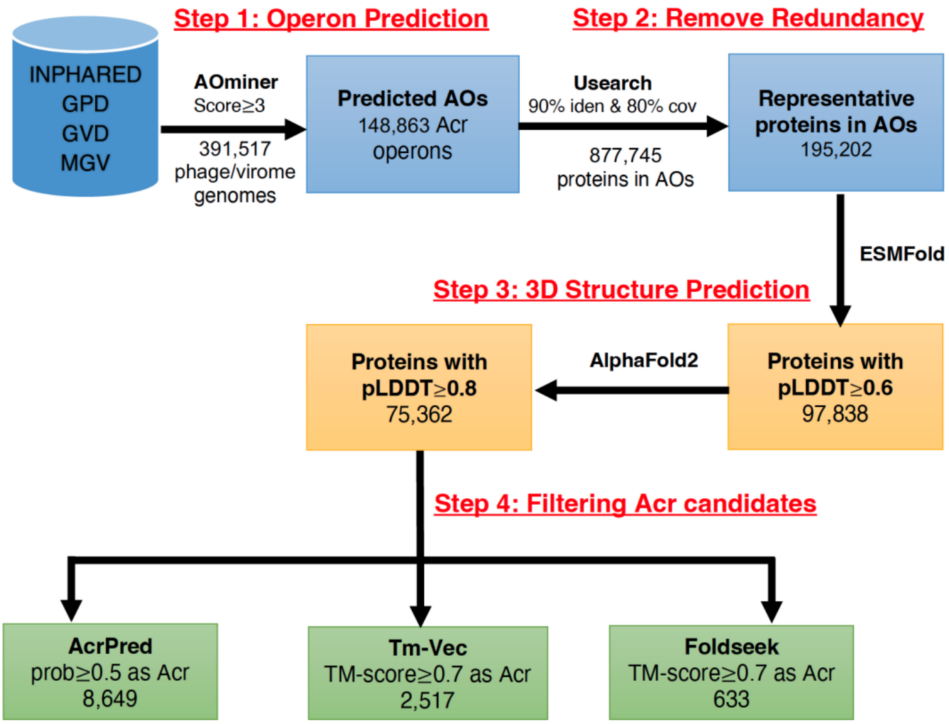

FIGURE 1 Workflow for identifying novel anti-CRISPRs from AOs using predicted 3D structures. AOs, anti-CRISPR operons; pLDDT, predicted local-distance difference test score; TM-score, template modeling score.

gene prediction in phage genomes. AOminer outputs a log of probability scores for each predicted operon. Operons with scores above 3 were kept as AOs.

2.3 | Remove sequence redundancy of proteins encoded in Acr operons

Once we get predicted AOs from AOminer, we clustered the proteins within these operons using Usearch (Edgar, 2010) with ≥90% sequence identity and ≥80% coverage, resulting in a non-redundant protein set. Members of each Usearch cluster were saved in a file for later use.

2.4 | Structure prediction

Representative proteins from the non-redundant set were first subjected to 3D structure prediction using ESMfold. It is 60 times faster than AlphaFold2 (Lin et al., 2023), allowing us to effectively process a large number of proteins. ESMfold and AlphaFold2 output a predicted local-distance difference test (pLDDT) score (a measure of per-residue accuracy of the predicted structure). This score ranges from 0 to 1, with higher scores indicating greater confidence in the predicted structure. ESMfold predicted structures with a pLDDT score ≥0.6 were selected for further refinement with AlphaFold2 to improve accuracy for those structure predictions. Finally, all structures from AlphaFold2 with a pLDDT score ≥0.8 were selected as high-confidence structures for further analysis. Recent papers (Park et al., 2022; Sahakyan et al., 2023) compared AlphaFold2 predicted Acr structures and their experimentally determined structures. They found that a pLDDT score ≥0.8 corresponds to minor differences between the predicted Acr structure and its real structure, with an RMSD (root mean squared deviation) between Cα atom coordinates <2 Å. This means predicted Acr structures with pLDDT scores ≥0.8 are of high confidence.

2.5 | Acr candidate prediction

Once we obtained the protein 3D structures, we used three tools to further analyze them for structural similarity with 122 experimentally characterized Acrs (Bondy-Denomy et al., 2018) or for sequence feature similarity using a machine learning approach:

1. *TM-Vec*: A structure-aware sequence–sequence comparison tool that embeds protein sequences using protein language models (Hamamsy et al., 2023). It searches for embedding similarity to indicate structural similarity between two sequences even when sequence identity is low. TM-Vec was used to compare predicted protein sequences (query) against known Acr sequences (target). It

returned the top five known Acr hits with a TM-score normalized for both query and target. Queries with TM-scores ≥0.7 were considered as high-confidence Acr candidates. This threshold was chosen as it is more stringent than a recent study, which used a less stringent threshold (TM-score ≥0.6) to select Acr structural homologs (Duan et al., 2024). More importantly, a TM-score ≥0.7 corresponds to a posterior probability >90% for two structurally aligned structures belonging to the same structural family of SCOP (Andreeva et al., 2008) or CATH (Dawson et al., 2017), according to a carefully conducted benchmark analysis (Xu & Zhang, 2010). This threshold is also supported by a more recent benchmark study (Wu et al., 2020) using CASP (Critical Assessment of Structure Prediction) data.

2. *Foldseek*: A structure-to-structure comparison tool that identifies similarities between protein structures (Van Kempen et al., 2024). We used Foldseek to compare predicted structures (query) against known Acr structures (target) to find structural homologs. For each query and target pair, Foldseek returned an Evalue, which is a TM-score normalized for both query and target. Queries with Foldseek Evalue ≥0.7 or higher were considered high-confident Acr candidates. The Evalue is calculated in Foldseek as a normalized TM-score between query and subject: (qTMscore + tTMscore)/2. This Evalue ≥0.7 threshold is chosen to be consistent with the above TM-Vec threshold.
3. *AcrPred*: A machine learning-based tool that uses sequence and evolutionary features to predict Acr proteins (Dao et al., 2023). Predicted protein sequences were run through AcrPred, which provides a probability score indicating the likelihood of a protein being an Acr. Proteins with a probability score ≥0.5 were classified as Acr candidates. Binary classification models like AcrPred output a probability score between 0 and 1. A score ≥0.5 represents the model's decision boundary, indicating the protein is more likely to be an Acr than a non-Acr.

2.6 | Mapping Acr candidates to Acr operons

We identified proteins supported by at least two of the three tools (Foldseek, TM-Vec, and AcrPred) as Acr candidates. These candidates were then mapped back to the original AOs (Figure 1, Step 1) identified by AOminer. In other words, only AOs containing at least one Acr candidate are included in the updated AcrDB.

To investigate if these AOs also contain sequence homologs of known Acrs, we used PSI-BLAST (Altschul et al., 1997) to search all proteins (query) encoded in these AOs against 122 known Acrs (target). Proteins with an *E*-value <0.001 and target coverage ≥80% were considered Acr homologs with sequence similarity to known Acrs. The purpose of this step was to identify AOs that not only contain new Acr candidates but also known Acr homologs. Graphical visualization of selected Acr operons was plotted using clinker (Gilchrist & Chooi, 2021).

2.7 | Clustering Acr candidates into Acr families

Proteins supported by all three tools (Foldseek, TM-Vec, and AcrPred) were identified as the most high-confident Acr candidates, representing the intersection of predictions from these tools. We clustered sequences of high-confident Acr candidates and known Acrs using MMseqs2 (Steinegger & Söding, 2017) (>40% sequence identity and >80% alignment coverage) to generate Acr sequence families.

2.8 | Hierarchical clustering of Acr families based on structural similarity

Furthermore, we calculated the average TM-score (template modeling score, a measure of similarity between two structures) for each pair of Acr families. For example, Acr family 1 (AF1) has *m* structures, and Acr family 2 (AF2) has *n* structures. The TM-score for AF1–AF2 is averaged from $m \times n$ TM-scores calculated by Foldseek. Then, hierarchical clustering was applied to all families based on their pair-wise TM-scores and plotted as a heatmap. The hierarchical clustering dendrogram is saved in a newick format and used for iTOL visualization (Letunic & Bork, 2024).

2.9 | Acr family co-localization network in operons

We further mapped proteins in Acr families into AOs and their source virome/phage MAGs/genomes. We retrieved the prokaryotic hosts from the metadata files of four virome and phage genome databases. We plotted the taxonomic distribution of the prokaryotic hosts of phage genomes that contain the AOs in iTOL together with the clustering dendrogram.

From the mapping result of Acr families to AOs, we identified all Acr family pairs if they co-occur in at least one AO. The co-occurrence data is stored in a CSV file, with the first two columns containing Acr IDs and the third column representing the co-occurrence frequencies between pairs of Acrs. Using the aMatReader app, the CSV file was loaded into Cytoscape (Franz et al., 2023) to generate the network. Nodes represent Acr IDs and edges represent co-occurrence relationships. Self-loops (diagonal values) are excluded during

the network creation. In the Style tab, the edge thickness was adjusted to reflect co-occurrence frequency.

3 | RESULTS

3.1 | 795 Acr candidate structures with 121 supported by three tools

From 391,517 metagenome-assembled genomes (MAGs) and isolate genomes of four phage/virome databases, AOminer predicted 148,863 Acr operons (Figure 1). These operons encode in total 877,745 proteins. After removing sequence redundancy, 195,202 representative protein sequences remained. Among these proteins, 75,362 proteins had highly confident AlphaFold2 predicted protein structures (pLDDT score ≥0.8). These proteins were subject to filters by three different tools for their similarity to 122 known Acrs (Bondy-Denomy et al., 2018) in terms of structures (Foldseek and TM-Vec) and sequence feature embeddings (AcrPred). The results of the three protein filtering processes are summarized in Table 2. In total, 795 out of the 75,362 proteins are predicted as Acr candidates, as they are supported by at least two of the three tools (AcrPred, TM-Vec, and Foldseek, Figure 2a). Structures predicted by a single tool were removed from this article and our AcrDB website to minimize false positives.

We further looked at the sequence similarity of the 795 Acrs to 122 known Acrs using PSI-BLAST (*E*-value <0.001 and target coverage >80%). Notably, 203 (25.5%) of the 795 had both sequence similarity and structure similarity to known Acrs. Among the 121 most confident Acrs (supported by all three tools, Figure 2a), 47 (38.8%) were also supported by PSI-BLAST. Given that most Acr candidates do not share sequence similarity with known Acrs, this confirms that Foldseek, TM-Vec, and AcrPred are more sensitive than traditional sequence similarity-based methods in detecting distant homology to known Acrs.

TABLE 2 Summary of high-confident AlphaFold2 structures for predicted Acrs.

Database	# of structures with pLDDT ≥0.8	AcrPred	Tm-Vec	Foldseek
INPHARED	6155	1072	235	72
GVD	5765	614	200	59
MGV	16,560	1979	525	158
GPD	46,882	4984	1557	344
Total	75,362	8649	2517	633

3.2 | Structure similarity-based hierarchical clustering of 163 Acr families

Combining 121 most confident candidate Acrs and 122 known Acrs, we clustered them (243 in total, >90% with pLDDT score ≥0.8) into potential Acr families (Figure 2b). MMseqs2 was used with a >40% sequence identity threshold and >80% coverage. This clustering process produced a total of 163 Acr families, among which 24 families contain two or more members and 139 families are singletons (including 105 known Acrs). These 24 non-singleton families contain in total 104 proteins, among which 14 families contain 17 known Acrs. Therefore, among the 163 Acr families, 119 are known Acr families, which received family labels of the contained longest known Acrs (e.g., AcrIIA5). For the 44 new Acr families, we named them "cAcr1" to "cAcr44," where "c" means "candidate" (Figure 2b). We used the 163 Acr families (24 non-singleton

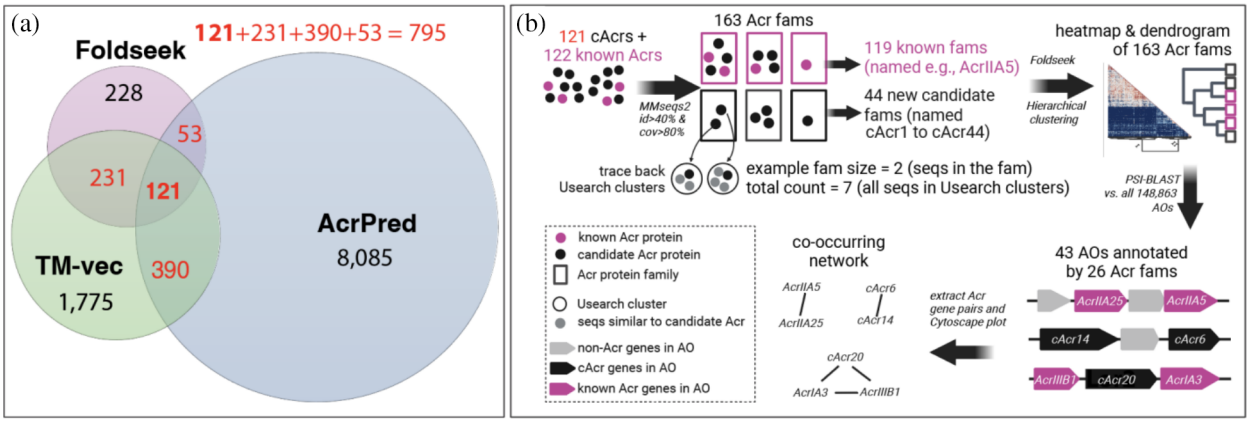

FIGURE 2 Data content and workflow for the detailed analysis of high-confident Acr candidates. (a) Venn diagram illustrating the overlaps among three tools (AcrPred, TM-Vec, Foldseek). The Acr candidate count in the intersection is 121. Candidates supported by at least two tools are 795. (b) Workflow to analyze the 121 most high-confident Acr candidates and the 122 known Acrs. Usearch clusters are explained in Step 2 of Figure 1, and sequences in the clusters are >90% identical to the 121 Acr candidates. AOs are explained in Step 1 of Figure 1. Other analyses are explained in the main text.

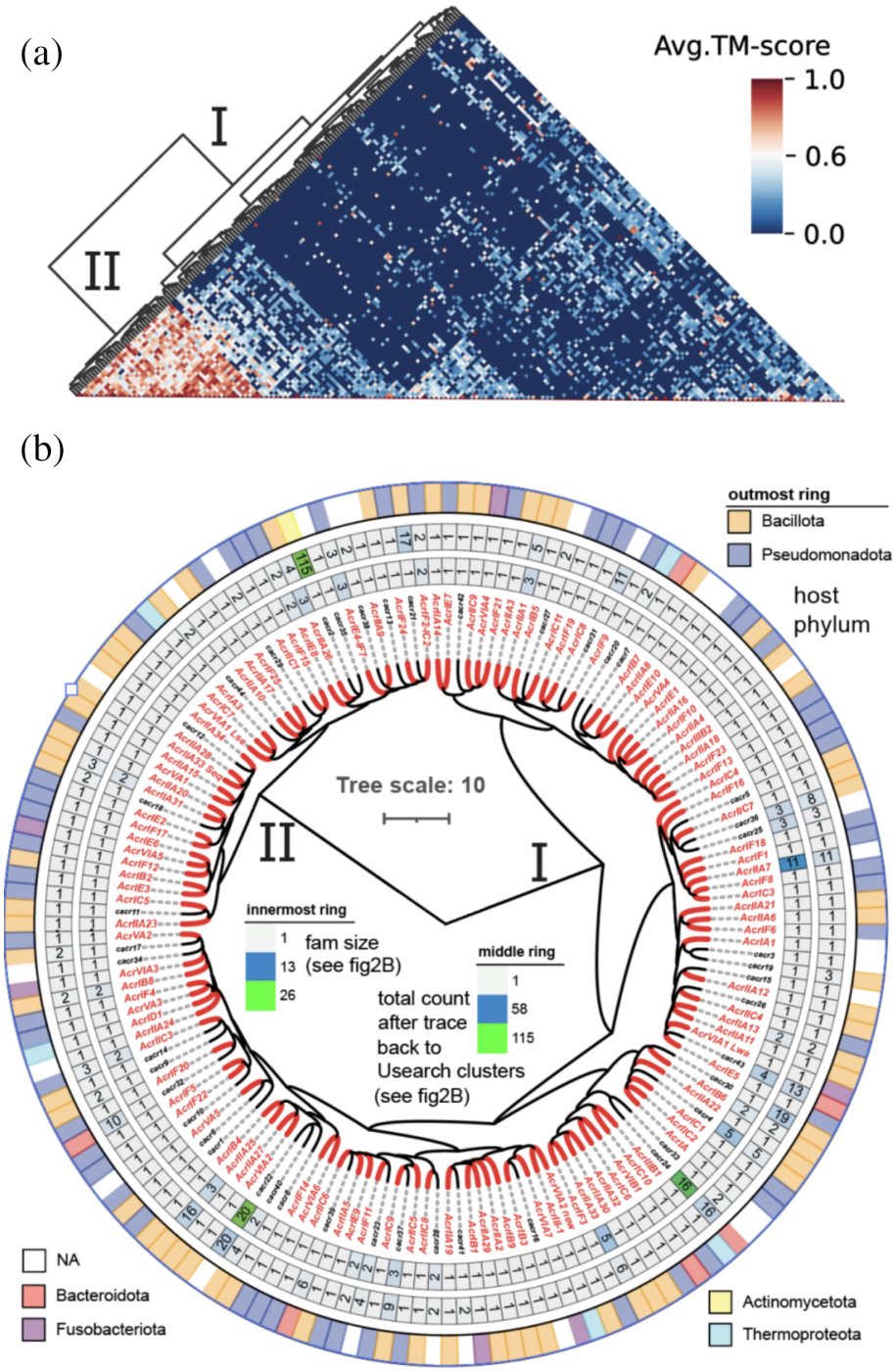

FIGURE 3 Structure similarity-based hierarchical clustering of 163 Acr families. (a) Heatmap of hierarchical clustering using the average TM-score for each pair of Acr families (see Methods). The dendrogram is also shown as the result of hierarchical clustering. (b) The dendrogram is visualized using iTOL. The outmost ring shows the host taxonomic phyla, the innermost ring shows the family size (Figure 2b), and the middle ring shows the total protein count considering all sequences in Usearch clusters (Figure 2b).

families, 34 cAcr singletons and 105 known Acr singletons) for the hierarchical clustering based on their structural similarities.

The 163 Acr families have <40% inter-family sequence identities, which was confirmed by a separate global sequence alignment using the needle program of EMBOSS (Rice et al., 2000). To study how they are structurally related, we performed a hierarchical clustering based on their structural similarities (see Methods). The clustering is presented as a heatmap in Figure 3a, showing two major clades (Figure 3a). Clade I contains 123 families, including 34 cAcr families and 89 known Acr families, with an average TM-score = 0.101. Clade II contains 40 families, including

10 cAcr families and 30 known Acr families, with an average TM-score = 0.648.

To better visualize the dendrogram from the hierarchical clustering, we plotted it using iTOL together with multiple rings to show the family member counts and their bacterial host taxonomy (Figure 3b). The 119 known Acr families are spread across the tree. This confirms that the 122 known Acr proteins are of rather distant families (<40% sequence identity) and strengthens the point that a large sequence diversity of Acrs in nature awaits discovery. Similarly, the 44 cAcrs are also distributed widely in the tree, and in many cases clustered closely with known Acrs, suggesting high structural similarities and potentially similar functions.

3.3 | Size and host taxonomic distribution of 163 Acr families

Each leaf in the tree represents a protein family (Figure 2b). Some Acr families were notably large, with certain families containing over 20 members (innermost ring of Figure 3b). For all the members in the family, we further traced back to their original Usearch clusters (Figure 1, Step 2) and counted all the sequences (Figure 2b) to obtain the total count of proteins represented by the family representative sequence. Now the largest family contains 115 proteins. Altogether, the 163 Acr families represent in total 470 sequences from the four source virome/phage databases and the 122 known Acrs (Table 3).

From the 163 family representative sequences, we extracted their prokaryotic host information and taxonomic lineages (26 have unknown hosts). Five bacteria phyla and one archaea phylum (*Thermoproteota*) are present in the tree (Figure 3b). The top phyla (Table 3) with the most Acr families and sequences are *Bacillota* (63 families and 127 sequences), *Pseudomonadota* (56 families and 236 sequences), and *Bacteroidota* (6 families and 13 sequences). This is expected as they are the major bacteria phyla found in the human gut microbiome.

3.4 | Co-localization network of 163 Acr families in Acr operons

Using the 163 Acr families, we further annotated the original 148,863 AOs (Figure 1) identified by AOminer. We used PSI-BLAST to search the 877,745 proteins encoded in these operons against the representative sequences of 163 Acr families (Figure 2b). We filtered the search result using *E*-value <0.001, target coverage >80%, and alignment identity >40%. Query proteins meeting these criteria received the Acr family labels of their best BLAST hits.

Out of the 877,745 proteins, 1415 were successfully annotated. These proteins belonged to 1361 operons, of which 43 operons contained at least two annotated Acr proteins. The proteins within these 43 operons were assigned to 26 of the 163 Acr families (Figure 2b). Within these 43 operons, 35 operons contained two Acr genes, 5 operons contained three Acr genes, and 3 operons contained four genes. Then we counted the Acr family–family co-occurrence in the 43 operons and visualized the results using Cytoscape.

This network plot, shown in Figure 4a, represents the co-occurrence relationships among Acr families identified in AOs. The largest cluster of co-localized Acrs contains 12 known Acr families (shown in pink), which are all AcrIIA families except AcrIB6. An example co-localization plot is shown for three AcrIIA families in four homologous operons (Figure 4b). AcrIIA25 is present in all four operons, while AcrIIA31 and AcrIIA5 are in two operons. These operons are from two source databases: INPHARED (*Streptococcus* phages) and MGV. This supports the fact that different Acrs and other anti-defense genes can recombine in different phages to form different operon structures as anti-defense islands (Pinilla-Redondo et al., 2020; Samuel et al., 2024). Three additional co-localization operon plots are provided in Figure S1, where more co-occurred AcrIIA genes in Figure 4a are illustrated. In Figure 4a, one cluster of co-localized Acrs contains both known Acrs (AcrIA3 and AcrIIIB1) and a candidate Acr (cAcr20). These three Acr families are co-localized in two operons (Figure 4c) from the INPHARED

TABLE 3 Summary of host taxonomic distribution of 163 Acr families and member proteins.

Phylum name/synonym	Acr families	Family %	Acrs in families	Total counts
Bacillota/*Firmicutes*	63	38.65	85	127
Pseudomonadota/*Proteobacteria*	56	34.36	105	236
Bacteroidota/*Bacteroidetes*	6	3.68	8	13
Fusobacteriota/*Fusobacteria*	6	3.68	6	21
Thermoproteota/*Crenarchaea*	5	3.07	5	5
Actinomycetota/*Actinobacteria*	1	0.61	1	1
Unknown hosts	26	15.95	33	67
Total	163	100	243	470

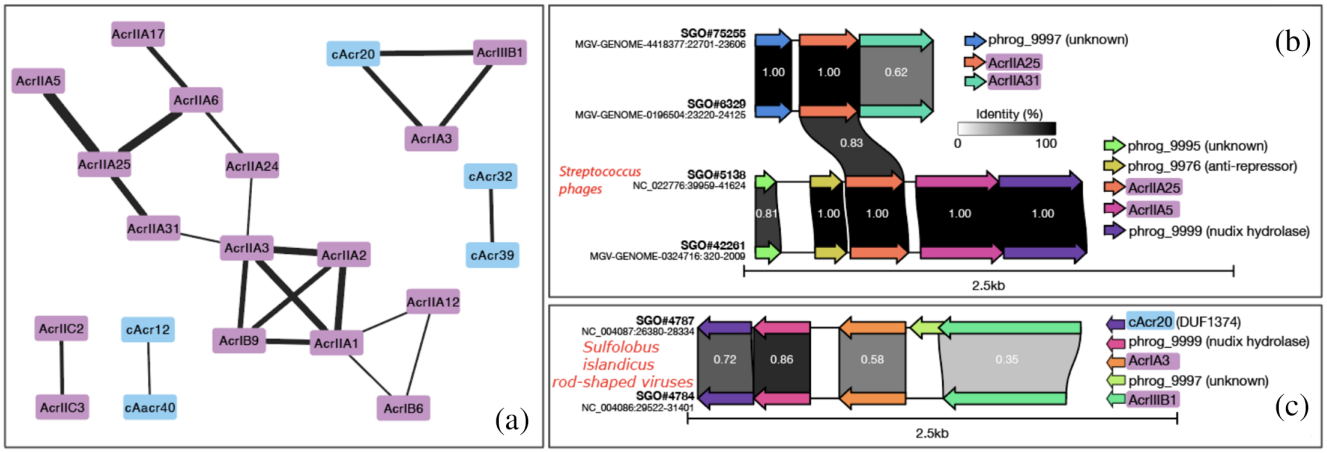

FIGURE 4 Co-occurrence network of Acr families in 43 AO operons. (a) Only operons with at least two different Acr families are used to generate this network plot, and operons with two Acr genes of the same family are excluded. Nodes represent Acr families and edges represent their co-occurrence within the same operons. Edge thickness reflects the co-occurrence frequency. (b) An example of the occurrence of known AcrIIAs in four operons. (c) An example of the occurrence of cAcr20, AcrIA3, and AcrIIIB1 in three operons. These plots are made by clinker (Gilchrist & Chooi, 2021). The operon (SGO, short gene operon) ID and source database ID, and operon genomic locations are indicated on the left. Arrows are genes and annotated gene labels are given on the right. PHROG (Terzian et al., 2021) is used to annotate non-Acr genes, and PHROG family IDs are given. Between operons, links are shown in different colors indicating the sequence identity between orthologous genes.

database (*Sulfolobus islandicus* rod-shaped viruses). The three Acrs are present in the two operons in the same order while the sequence identity varies. The fact that cAcr20 (with Pfam DUF1374 domain) is co-localized with known Acrs suggests that it is likely also involved in anti-CRISPR functions.

3.5 | Website organization and structural similarity search web service

Our first version of AcrDB is available at https://bcb.unl.edu/AcrDB/. For this updated database, we provide a mirror website at https://pro.unl.edu/AcrDB and modify it to include the herein reported new data and functions. From the homepage, users can access Acr operon data from the four source databases. For example, clicking on the INPHARED link (inset of Figure 5f) will open a new page with a table (Figure 5a,b). Clicking on an operon ID will open the operon page, where users can view the operon structure as a graph (Figure 5c). The operon page also has five tabs (Figure 5d) including a structure tab with the predicted 3D structure visualization and tabs for results from the three Acr prediction tools (e.g., Figure 5d for TM-Vec and Figure 5e for Foldseek).

From anywhere on the website, users can navigate through the top navigation menu (Figure 5f) to access the Download page, Help page, and AOminer GitHub page. Importantly, we also provide the "Structure Similarity Search" page, where users can submit their own sequence or PDB structure to search against the protein structures of 122 known Acrs using our backend computing server. Additionally, for the 122 known Acrs ("Known_Acrs" link of the inset of Figure 5f), we provide a webpage for users to easily access all the necessary information including their targeted CRISPR-Cas subtypes, species of origin, literature, sequences, and 3D structures. We also provide a link where users can access the interactive heatmaps of the structure-based hierarchical clustering of 163 Acr families (Figure 3a) and of the 122 known Acrs.

4 | DISCUSSION

In this study, we focused our new Acr discovery in human gut viromes. Human gut virome databases such as GPD (Camarillo-Guerrero et al., 2021), GVD (Gregory et al., 2020), MGV (Nayfach et al., 2021) are highly valuable for new anti-CRISPR discovery. Most viral genomes in these databases are contigs assembled from metagenomic reads. The diversity and abundance of phages in these gut viromes increase the likelihood of finding novel Acr proteins. Recently, we have computationally screened human gut viromes for novel *Acr-Aca* operons and other anti-prokaryotic immune system genes (Yan et al., 2023; Yang et al., 2022; Yang et al., 2023). The human gut microbiome and virome have also experimentally proven to be a source for Acr discovery. For example, four Acr proteins, AcrIIA7, AcrIIA8, AcrIIA9, and AcrIIA10 were identified from metagenomic libraries of the human gut. These proteins can inhibit the type II-A CRISPR-Cas system utilized by SpyCas9 (Forsberg et al., 2019).

Compared to the first version of AcrDB, we replaced AcrFinder (Yi et al., 2020) with AOminer (Yang et al., 2023) for identifying operons that potentially encode Acrs. The algorithm of AcrFinder requires the presence of *Aca* genes in the operons, which does not

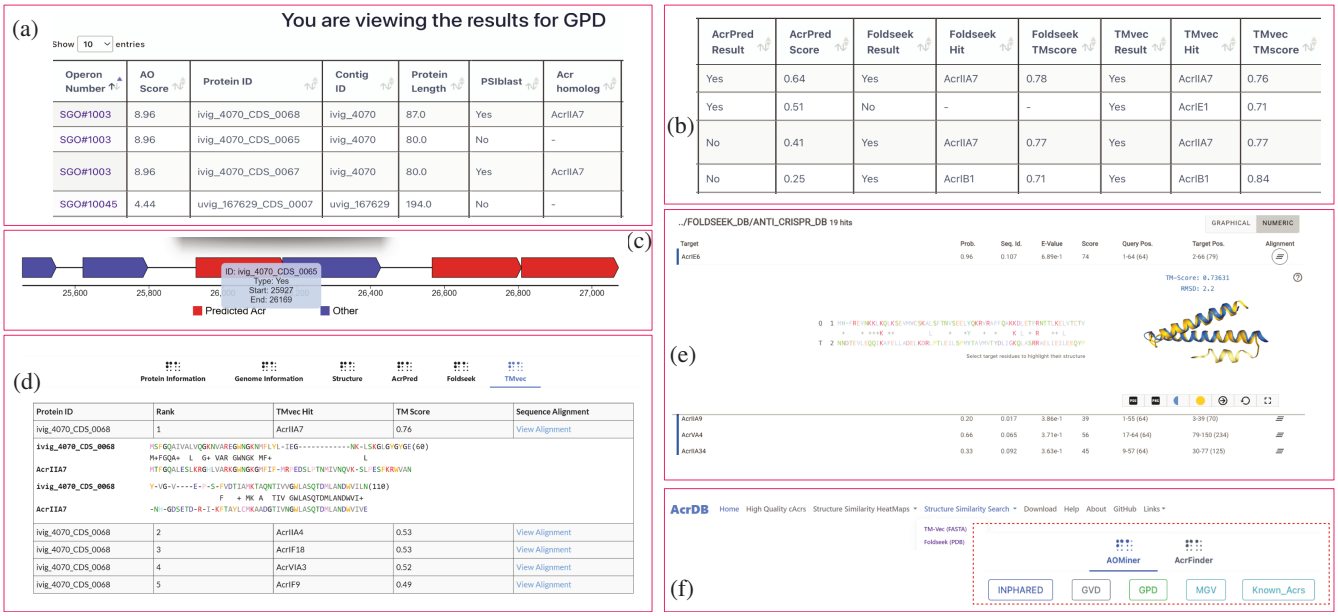

FIGURE 5 Screenshots of AcrDB website. (a) AOs predicted from INPHARED phage genomes. Only AOs that contain at least one candidate Acrs are included in the table. (b) The table has many columns, which cannot fit in one screen. A scroll bar is provided so that users can see columns on the far right side. (c) Clicking on the SGO ID will open the operon page, where on the top is the graphical view of the operon structure. Mouse-over the red genes (Acrs) will show the gene ID and position information. (d) Below the operon structure graph in the operon page are six tabs for each protein encoded in the operon. Clicking on each gene in the operon structure graph will automatically update the contents in the six tabs. As an example, the TM-Vec tab has a table to show the top five known Acr hits of the third red gene in c. (e) Clicking on the Foldseek tab will show a link to the current page, which is the Foldseek search result against known Acrs. (f) Inset: from AcrDB homepage, users can choose to access data in the old version of AcrDB by clicking on the AcrFinder tab. By default, the AOminer tab is on. From the navigation menu on the top, users can click on "Structure Similarity Search" to access the TM-Vec and Foldseek search pages.

apply to Acrs without the need of Acas. AOminer does not have this limitation, as its algorithm can distinguish between operons containing putative Acrs and non-Acr operons, allowing researchers to focus on genomic regions with a higher likelihood of Acr presence. These predicted operons can then be further investigated with other bioinformatics tools or experimental approaches to discover novel Acrs.

This updated AcrDB features the use of structural similarity tools for new Acr discovery. This idea has been highlighted in a review paper (Makarova et al., 2023) published in 2023, and successfully employed to assist the experimental characterization of new Acrs from viromes (Duan et al., 2024) in a paper published in 2024. It should be noted that structural similarities do not always mean that our candidate Acrs share the same evolutionary origins with their respective best known Acr hits. Most likely, there are multiple evolutionary routes for the origin of Acrs, such as repurposing host-derived proteins (e.g., examples in Sahakyan et al., 2023) followed by rapid sequence divergence, convergent evolution from unrelated proteins, or de novo from non-genic DNA, and so forth. Two structural similarity search tools are used to find Acr candidates from Acr operons. TM-Vec (Hamamsy et al., 2023) identified a significantly higher number of Acr candidates than Foldseek (Van Kempen et al., 2024) (Table 2). Even though both tools detect structural similarity, Foldseek uses a structural alphabet-based method for structural alignment between two 3D structures and directly detects structural similarity. In contrast, TM-Vec indirectly detects structural similarity as it uses structure-aware sequence embedding similarity between two proteins to indicate structural similarity. Future studies will be needed to compare Foldseek and TM-Vec in terms of their accuracy.

Overall, our result emphasizes the importance of using multiple structure-based tools for a comprehensive Acr candidate identification. By integrating these tools, we ensured a thorough analysis of the predicted protein structures, identifying candidates with both structural and sequence-based similarities to known Acrs. This comprehensive, multi-tool approach allowed us to prioritize the most promising Acr candidates for further experimental validation.

Two recent studies predicted 3D structures for experimentally characterized Acrs (Park et al., 2022; Sahakyan et al., 2023). However, their data are not available on a website that supports easy search, query, and visualization of the structures. Also, no online databases are available for 3D structures of predicted Acrs in viromes. Our updated AcrDB addressed all these issues and provides a web server for users to

submit their own sequences and structures to search against structures of 122 known Acrs. AcrDB is designed to serve as a foundational resource for the anti-CRISPR research community, providing user-friendly access to computationally predicted Acr structures in the context of Acr operons.

AUTHOR CONTRIBUTIONS

Minal Khatri: Methodology; software; data curation; investigation; validation; formal analysis; visualization; writing – original draft. **N. R. Siva Shanmugam:** Methodology; data curation; software; validation; investigation; formal analysis; visualization; writing – review and editing. **Xinpeng Zhang:** Methodology; investigation; formal analysis; visualization. **Revanth Sai Kumar Reddy Patel:** Methodology; software; data curation; formal analysis; visualization. **Yanbin Yin:** Conceptualization; methodology; data curation; investigation; validation; supervision; funding acquisition; visualization; project administration; resources; writing – review and editing.

ACKNOWLEDGMENTS

This work was partially completed utilizing the Holland Computing Center of the University of Nebraska-Lincoln, which receives support from the Nebraska Research Initiative.

FUNDING INFORMATION

This work was supported by the U.S. National Institutes of Health (NIH) awards (R01GM140370 and R21AI171952) and U.S. Department of Agriculture (USDA) award (58-8042-9-089).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in AcrDB at https://pro.unl.edu/AcrDB/index.php.

ORCID

Yanbin Yin https://orcid.org/0000-0001-7667-881X

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, et al. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008;36:D419–25.

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373:871–6.

Birkholz N, Kamata K, Feussner M, Wilkinson ME, Cuba Samaniego C, Migur A, et al. Phage anti-CRISPR control by an RNA- and DNA-binding helix-turn-helix protein. Nature. 2024;631:670–7.

Bondy-Denomy J. Protein inhibitors of CRISPR-Cas9. ACS Chem Biol. 2018;13:417–23.

Bondy-Denomy J, Davidson AR, Doudna JA, Fineran PC, Maxwell KL, Moineau S, et al. A unified resource for tracking anti-CRISPR names. CRISPR J. 2018;1:304–5.

Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature. 2013;493:429–32.

Borges AL, Davidson AR, Bondy-Denomy J. The discovery, mechanisms, and evolutionary impact of anti-CRISPRs. Annu Rev Virol. 2017;4:37–59.

Bouras G, Nepal R, Houtak G, Psaltis AJ, Wormald P-J, Vreugde S. Pharokka: a fast scalable bacteriophage annotation tool. Bioinformatics. 2023;39:btac776.

Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell. 2021;184:1098–109.

Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, et al. INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. Phage. 2021;2:214–23.

Dao F-Y, Liu M-L, Su W, Lv H, Zhang Z-Y, Lin H, et al. AcrPred: a hybrid optimization with enumerated machine learning algorithm to predict anti-CRISPR proteins. Int J Biol Macromol. 2023;228:706–14.

Davidson AR, Lu W-T, Stanley SY, Wang J, Mejdani M, Trost CN, et al. Anti-CRISPRs: protein inhibitors of CRISPR-Cas systems. Annu Rev Biochem. 2020;89:309–32.

Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, et al. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2017;45:D289–95.

Dong C, Wang X, Ma C, Zeng Z, Pu D-K, Liu S, et al. Anti-CRISPRdb v2. 2: an online repository of anti-CRISPR proteins including information on inhibitory mechanisms, activities and neighbors of curated anti-CRISPR proteins. Database. 2022;2022:baac010.

Duan N, Hand E, Pheko M, Sharma S, Emiola A. Structure-guided discovery of anti-CRISPR and anti-phage defense proteins. Nat Commun. 2024;15:649.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.

Eitzinger S, Asif A, Watters KE, Iavarone AT, Knott GJ, Doudna JA, et al. Machine learning predicts new anti-CRISPR proteins. Nucleic Acids Res. 2020;48:4698–708.

Forsberg KJ, Bhatt IV, Schmidtke DT, Javanmardi K, Dillard KE, Stoddard BL, et al. Functional metagenomics-guided discovery of potent Cas9 inhibitors in the human microbiome. Elife. 2019;8:e46540. https://doi.org/10.7554/eLife.46540

Franz M, Lopes CT, Fong D, Kucera M, Cheung M, Siper MC, et al. Cytoscape.js 2023 update: a graph theory library for visualization and analysis. Bioinformatics. 2023;39(1):btad031.

Gilchrist CLM, Chooi YH. Clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics. 2021;37:2473–5.

Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe. 2020;28:724–40.

Gussow AB, Park AE, Borges AL, Shmakov SA, Makarova KS, Wolf YI, et al. Machine-learning approach expands the repertoire of anti-CRISPR protein families. Nat Commun. 2020;11:3784.

Hamamsy T, Morton JT, Blackwell R, Berenberg D, Carriero N, Gligorijevic V, et al. Protein remote homology detection and structural alignment using deep learning. Nat Biotechnol. 2023;42(6):975–85. https://doi.org/10.1038/s41587-023-01917-2

Huang L, Yang B, Yi H, Asif A, Wang J, Lithgow T, et al. AcrDB: a database of anti-CRISPR operons in prokaryotes and viruses. Nucleic Acids Res. 2021;49:D622–9.
Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinforma. 2010;11:1–11.
Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2
Letunic I, Bork P. Interactive tree of life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res. 2024;52:W78–82.
Li Y, Wei Y, Xu S, Tan Q, Zong L, Wang J, et al. AcrNET: predicting anti-CRISPR with deep learning. Bioinformatics. 2023;39:btad259.
Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379:1123–30.
Makarova KS, Wolf YI, Koonin EV. In silico approaches for prediction of anti-CRISPR proteins. J Mol Biol. 2023;435:168036.
Mayo-Muñoz D, Pinilla-Redondo R, Camara-Wilpert S, Birkholz N, Fineran PC. Inhibitors of bacterial immune systems: discovery, mechanisms and applications. Nat Rev Genet. 2024;25:237–54.
Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol. 2021;6:960–70.
Park H-M, Park Y, Vankerschaver J, Van Messem A, De Neve W, Shim H. Rethinking protein drug design with highly accurate structure prediction of anti-CRISPR proteins. Pharmaceuticals. 2022;15(3):310. https://doi.org/10.3390/ph15030310
Pawluk A, Davidson AR, Maxwell KL. Anti-CRISPR: discovery, mechanism and function. Nat Rev Microbiol. 2018;16:12–7.
Pinilla-Redondo R, Shehreen S, Marino ND, Fagerlund RD, Brown CM, Sørensen SJ, et al. Discovery of multiple anti-CRISPRs highlights anti-defense gene clustering in mobile genetic elements. Nat Commun. 2020;11:5652.
Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16:276–7.
Sahakyan H, Makarova KS, Koonin EV. Search for origins of anti-CRISPR proteins by structure comparison. CRISPR J. 2023;6:222–31.
Samuel B, Mittelman K, Croitoru SY, Ben Haim M, Burstein D. Diverse anti-defence systems are encoded in the leading region of plasmids. Nature. 2024;635:186–92.
Stanley SY, Maxwell KL. Phage-encoded anti-CRISPR defenses. Annu Rev Genet. 2018;52:445–64.
Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8.
Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio Rubén E, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genomics Bioinform. 2021;3(3):lqab067.
Tesson F, Huiting E, Wei L, Ren J, Johnson M, Planel R, et al. Exploring the diversity of anti-defense systems across prokaryotes, phages and mobile genetic elements. Nucleic Acids Res. 2025;53(1):gkae1171.
Van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CL, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42:243–6.
Wandera KG, Alkhnbashi OS, Bassett H, Mitrofanov A, Hauns S, Migur A, et al. Anti-CRISPR prediction using deep learning reveals an inhibitor of Cas13b nucleases. Mol Cell. 2022;82:2714–2726.e4.
Wang J, Dai W, Li J, Li Q, Xie R, Zhang Y, et al. AcrHub: an integrative hub for investigating, predicting and mapping anti-CRISPR proteins. Nucleic Acids Res. 2021;49:D630–8.
Wang J, Dai W, Li J, Xie R, Dunstan RA, Stubenrauch C, et al. PaCRISPR: a server for predicting and visualizing anti-CRISPR proteins. Nucleic Acids Res. 2020;48:W348–57.
Wu L-y, Xia X-y, Pan X-m. A novel score for highly accurate and efficient prediction of native protein structures. 2020. bioRxiv: 2020.2004.2023.056945.
Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010;26:889–95.
Yan Y, Zheng J, Zhang X, Yin Y. dbAPIS: a database of anti-prokaryotic immune system genes. Nucleic Acids Res. 2023;52:D419–25.
Yang B, Khatri M, Zheng J, Deogun J, Yin Y. Genome mining for anti-CRISPR operons using machine learning. Bioinformatics. 2023;39:btad309.
Yang B, Zheng J, Yin Y. AcaFinder: genome mining for anti-CRISPR-associated genes. mSystems. 2022;7:e00817-00822.
Yi H, Huang L, Yang B, Gomez J, Zhang H, Yin Y. AcrFinder: genome mining anti-CRISPR operons in prokaryotes and their viruses. Nucleic Acids Res. 2020;48:W358–65.
Yin Y, Yang B, Entwistle S. Bioinformatics identification of anti-CRISPR loci by using homology, guilt-by-association, and CRISPR self-targeting spacer approaches. mSystems. 2019;4(5):e00455-19. https://doi.org/10.1128/msystems.00455-00419
Zhang F, Zhao S, Ren C, Zhu Y, Zhou H, Lai Y, et al. CRISPRminer is a knowledge base for exploring CRISPR-Cas systems in microbe and phage interactions. Commun Biol. 2018;1:180.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Khatri M, Shanmugam NRS, Zhang X, Patel RSKR, Yin Y. AcrDB update: Predicted 3D structures of anti-CRISPRs in human gut viromes. Protein Science. 2025;34(6):e70177. https://doi.org/10.1002/pro.70177